Infocommunications Journal

A PUBLICATION OF THE SCIENTIFIC ASSOCIATION FOR INFOCOMMUNICATIONS (HTE)

June 2025 Volume XVII Number 2 ISSN 2061-2079

Authors, co-authors of the June 2025 issue 1

PAPERS FROM OPEN CALL

Validation Methodology of Wireless Brain-Computer Interface for Event-Related Potential Application ..... *Ádám Salamon, Gábor Takács, and György Bognár* 2

Masked Face Image Inpainting Based on Generative Adversarial Network ..... *Qingyu Liu, Lei Chen, Yeguo Sun, and Lei Lius* 11

Monitoring the Semantic Change of COVID-19-related Expressions Using Dynamic Word Embeddings ..... *Bogdán Asztalos, and Péter Bányász* 20

A Comparative Analysis of Static Word Embeddings for Hungarian ..... *Máté Gedeon* 28

IP Packet Forwarding Performance Comparison of the FD.io VPP and the Linux Kernel ..... *Melinda Kosák, and Gábor Lencse* 35

Multipath Rate Control for Real-Time Media ..... *Balázs Kreith and Árpád Drozdy* 45

Completion Time Prediction of Open Source FaaS Functions ..... *David Balla, Markosz Maliosz, and Csaba Simon* 53

Framework for Intrusion Detection in IoT Networks: Dataset Design and Machine Learning Analysis ..... *Mansour Lmkaiti, Ibtissam Larhlimi, ... Maryem Lachgar, Houda Moudni and Hicham Mouncif* 61

Multi-photon QKD for Practical Quantum Networks ..... *ANitin Jha, Abhishek Parakh, and Mahadevan Subramaniam* 72

Mode Selection in Mode Division Multiple Access System for In Building Solution in Mobile Networks ..... *Ahmed S. Mohamed, and Eszter Udvary* 83

On the Enhancement Anomaly Detection for RF Bio-Sensors by Computing Artificial Networks Using Machine Learning Techniques ..... *Raya Adel Kamil, Saif Mohamed Baraa Alsabti, Rusul K. Abdulsattar, ... Ammar H. Mohammed, and Taha A. Elwi* 89

Constrained LS Channel Estimation for Massive MIMO Communication Systems ..... *Muhammad Ahsan Shaikh, Tayyab Ahmed Shaikh, ... Sadiq Ur Rehman, and Halar Mustafa* 96

CALL FOR PAPER / PARTICIPATION

EEE ICC 2026 / IEEE International Conference on Communications Glasgow, Scotland, UK 105

ADDITIONAL

Guidelines for our Authors 104

Technically Co-Sponsored by

IEEE ComSoc™ IEEE Communications Society

hte

IEEE HUNGARY SECTION

Indexing information

Infocommunications Journal is covered by Inspec, Compendex and Scopus.
Infocommunications Journal is also included in the Thomson Reuters – Web of ScienceTM Core Collection, Emerging Sources Citation Index (ESCI)

Infocommunications Journal

Technically co-sponsored by IEEE Communications Society and IEEE Hungary Section

Supporters

FERENC VÁGUJHELYI – president, Scientific Association for Infocommunications (HTE)

The publication was produced with the support of the Hungarian Academy of Sciences and the NMHH

Editorial Office (Subscription and Advertisements)**:**
Scientific Association for Infocommunications
H-1051 Budapest, Bajcsy-Zsilinszky str. 12, Room: 502
Phone: +36 1 353 1027 • E-mail: info@hte.hu • Web: www.hte.hu

Articles can be sent also to the following address:
Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
Phone: +36 1 463 4189 • E-mail: pvarga@tmit.bme.hu

Subscription rates for foreign subscribers: 4 issues 13.700 HUF + postage

Publisher: PÉTER NAGY

HU ISSN 2061-2079 • Layout: PLAZMA DS • Printed by: FOM Media

Authors, co-authors of the June 2025 issue

Ádám Salamon, Gábor Takács,
György Bognár, Qingyu Liu, Lei Chen,
Yeguo Sun, Lei Liu, Bogdán Asztalos,
Péter Bányász, Máté Gedeon,
Melinda Kosák, Gábor Lencse,
Balázs Kreith, Árpád Drozdy,
David Balla, Markosz Maliosz,
Csaba Simon, Mansour Lmkaiti,
Ibtissam Larhlimi, Maryem Lachgar,
Houda Moudni, Hicham Mouncif,
ANitin Jha, Abhishek Parakh,
Mahadevan Subramaniam,
Ahmed S. Mohamed, Eszter Udvary,
Raya Adel Kamil,
Saif Mohamed Baraa Alsabti,
Rusul K. Abdulsattar,
Ammar H. Mohammed, Taha A. Elwi,
Muhammad Ahsan Shaikh,
Tayyab Ahmed Shaikh,
Sadiq Ur Rehman, Halar Mustafa

Validation Methodology of Wireless Brain-Computer Interface for Event-Related Potential Application

Ádám Salamon, Gábor Takács, and György Bognár

***Abstract*—Electroencephalography (EEG) is a technique used to observe brain activity by measuring the dynamic changes of the electric field induced by neurons' activity. Brain-computer interface (BCI) systems are used in cognitive psychology examinations measuring the changes of brain activities. This paper presents a validation methodology to characterize BCI systems with wireless communication interface and the applicability on a preselected BCI system. This way, the delay, the functionality, and the frequency selectivity can be determined of the overall BCI system, taking into account the effect of the hardware, the software, and the electrodes, avoiding noise artifacts. The presented and validated BCI system proved to be successfully applied in ERP EEG measurements such as steady-state visually evoked potential, pattern-reversal visually evoked potential, and P300 event-related potential.**

***Index Terms*—Biomedical communication, Biomedical electrodes, Biomedical electronics, Biomedical engineering, Biomedical measurement, Biomedical signal processing, Brain modeling, Brain-computer interfaces, Electroencephalography.**

I. INTRODUCTION

Electroencephalography (EEG) is a commonly used technique to observe brain activity by measuring the dynamic changes of the electric field induced by the neurons. EEG provides excellent time domain and weaker spatial resolution of the brain activities and can show the functional state of the brain and its dynamic changes. Special brain-computer interface (BCI) systems are applied to different cognitive psychology examinations to measure the changing electrical activity of the brain. The validation of these BCI systems has to be carried out to determine the signal-to-noise ratio, the delay time, and exact applicability.

Although various heterogeneous validation techniques exist, only the American Clinical Neurophysiology Society (ACNS) provides standardized guidelines for the clinical use of EEG systems [1]. The methodology approaches presented in this paper are based on this guideline. However, the worked-out characterization methodology for BCI systems focuses on the most frequently applied event-related potential (ERP) measurements in cognitive psychology examinations (e.g., instead of P100, P300 was measured). The ERP is a non-invasive neuroimaging technique measuring the brain's electrical activity in response to a specific event or stimulus.

This paper presents the validation methodology for BCI systems with wireless communication interface toward a personal computer.

Ádám Salamon, Gábor Takács, and György Bognár are with Department of Electron Devices, Budapest University of Technology and Economics, Hungary (e-mail: salamon.adam@vik.bme.hu, takacs.gabor@vik.bme.hu, bognar.gyorgy@vik.bme.hu).

DOI: 10.36244/ICJ.2025.2.1

By the proposed methodology, not only the hardware part of the signal processing can be measured and validated, but also the effects of the software component(s), the electrodes, and the overall system. (Figure 1.)

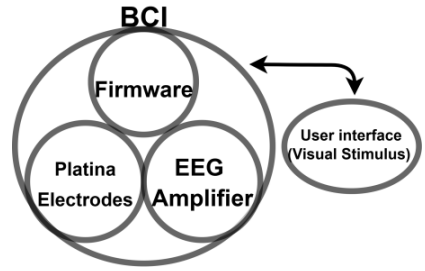


Figure 1.: BCI headset components and visual stimulus. [2, Fig. 1 adapted]

In this paper, the application of the developed characterization methodology on a selected BCI system (Figure 2.) intended for event-related potential (ERP) examinations is presented in detail.

Initially, a single photodiode was applied to simulate a simplified brain model. This approach allowed for precise characterization of the system's latency without measuring the additional noise introduced by the human brain and body. This setup made it possible to isolate and identify any intrinsic system-level deviations or artifacts that might affect signal integrity.

In the following steps, event-related potential measurements on multiple human subjects were carried out. This step aimed to assess the practical applicability and performance of the BCI system under real-world conditions, specifically evaluating whether the system fulfills the event-related potential (ERP) application requirements (can be seen in Table I.). These requirements typically include aspects such as signal clarity, timing accuracy, and the system's ability to detect event-related/evoked neural responses consistently across different individuals.

The presented methodology, combining both hardware-based testing and subject-based ERP experiments, demonstrated its applicability for validating BCI systems intended for ERP applications.

This approach ensures that technical performance and biological compatibility are thoroughly investigated. It also offers a robust framework for selecting and validating BCI platforms in non-invasive neuroimaging research and development.

Section II provides a concise overview of EEG fundamentals and different ERP measurement techniques applied in cognitive psychology examinations, which are summarized to give a solid background in this specific field of neuroscience.

Section III describes the validation methodology and the developed characterization system.
Section IV presents the selected BCI system in detail.
Section V discusses the characterization and measurement results.

Figure 2.: BCI headset prototype (flexible headset, measured area, electrodes).

II. THEORETICAL BACKGROUND

A. Electroencephalography

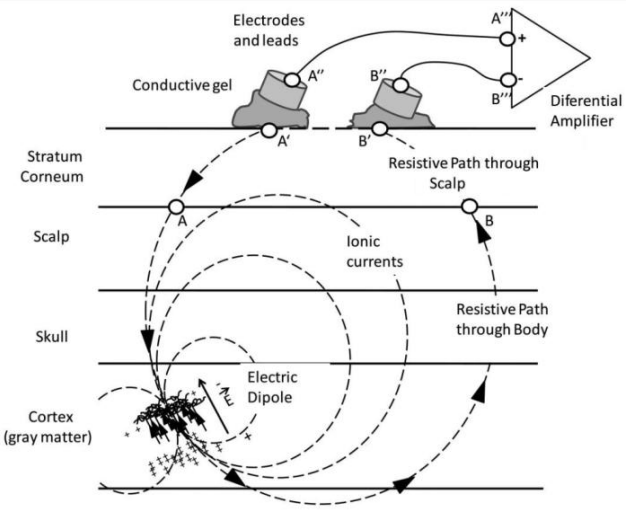


Figure 3.: Electric dipoles created by the postsynaptic potentials and EEG measurement. [3]

The extracranial EEG signals are observed on the surface of the hairy scalp, but several attenuating factors influence the signal (Figure 3.). Therefore, a synchronous potential change of at least 6-10 cm^2 of the cerebral cortex is required to achieve an evaluable signal-to-noise ratio [1], [2]. The attenuating factors significantly narrow the range of brain activities that can be examined and studied. With this limitation, EEG is still effectively used in neurological research and BCI applications [4], [5].

B. Neurons' activities

Neurons show two electrical activities: action potential and postsynaptic potential change. The action potential is triggered when the internal potential of the neuron reaches a value above a threshold level as a result of a stimulus from dendrites [6]. At this point, a self-sustaining (70–110 mV) potential change of the order of a millisecond extends from the cell to the ends of the axon [7], [8]. The time course and amplitude of the potential change are constant for the given cell. If the stimulus reaches the activation threshold, it is no longer independent of its parameters. In most cases, an action potential cannot be detected with electrodes placed on the scalp, except only in auditory-induced cerebral responses satisfying conditions for sensing the action potential change where several axons run in parallel.

In the EEG studies presented in this paper, the electric signals received with the electrodes are caused by postsynaptic potential changes in cortical pyramidal cells. Unlike the action potential, postsynaptic potential changes are graded potentials (100 µV to 10 mV) with slower (5 ms to 30 ms) duration [7]. The neuron performs the summation of the excitatory and inhibitory postsynaptic potentials. Excitatory postsynaptic potential brings the neuron closer to the action potential threshold, and inhibitory postsynaptic potentials move the neuron away from the action potential threshold [9].

Cortical pyramidal cells neurons are oriented parallel to each other, and the field electrical dipole generated by the postsynaptic potential can be measured with EEG through the scalp. (Figure 4.)

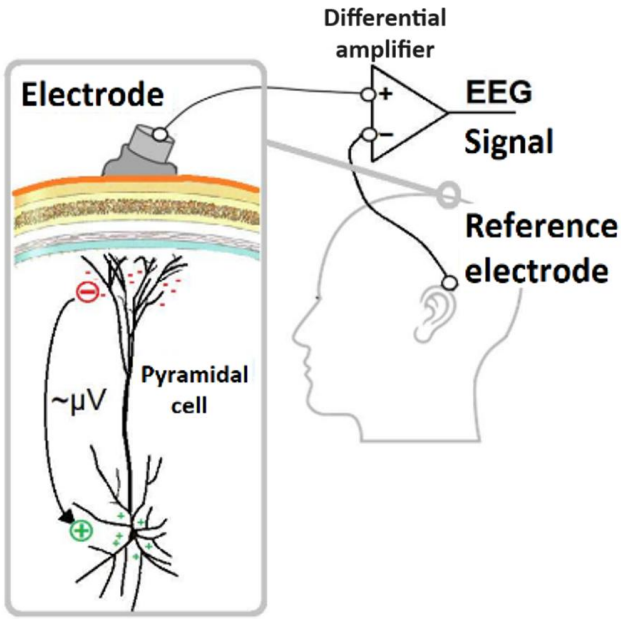


Figure 4.: Postsynaptic potential change in the pyramidal cell creates an electric dipole in the area around the neuron.

C. International 10-20 System for electrode placement

The 10-20 electrode placement system is the first and still accepted standard for defining and naming electrode positions. In this system, 21 electrodes are placed at 10% and 20% relative distances along the skull. (Figure 5.) Later, more electrode positions were added to the standard called 10-10 systems. Further extensions, like 5-10 systems and other electrode systems (that increased the number of electrodes) are used by

vendors but still not accepted as clinical EEG nomenclature standards [10].

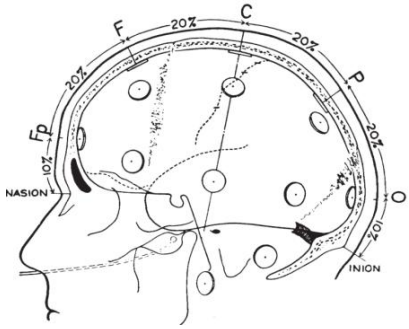


Figure 5.: International 10-20 System. [11].

D. Event-Related Potential

Event-related potential (ERP) is a direct response to a specific sensory, cognitive, or motor event of the brain [12]. BCI interfaces usually use these potential changes to determine user intentions.

The EEG signal represents many ongoing brain processes; therefore, a single event is invisible. Many trials are recorded and averaged to extract the specific brain response.

The signal-averaging procedure is used to extract event-Related Potentials from the EEG signal. This technique applies the following assumptions:

1. ERPs are invariable to signal latency and morphology.
2. The noise can be approximated by a zero-mean Gaussian random process uncorrelated between trials and not time-locked to the event.

Noise cannot be related to brain functions, muscle movements, and external electric fields. Averaging improves the signal-noise ratio with $\sqrt{N}$, where N is the number of epochs:

$$x(t) = ERP(t) + noise(k,t)$$

$$\bar{x}(t) = \frac{1}{N}\sum_{k=1}^{N} x(t,k) = ERP(t) + \frac{1}{N}\sum_{k=1}^{N} noise(t,k)$$

x represents an epoch of the recorded EEG during a trial, k is the epoch number, and *t* represents the time elapsed after the event.

The naming convention of event-related potentials follows the following rule:

- The first letter shows the potential change direction:
  - P – positive,
  - N – negative,
  - C – not defined.
- The following number defined the ordinal position of the peak in the waveform or the latency of the peak (e.g., P300 for a peak at 300 ms).

ERP component names can often be confused. Multiple ERPs can have the same name, but it can mostly be easily determined from the context [13].

E. P300 Event-Related Potential

P300 is a commonly used event-related potential (ERP) in BCI applications. Two P300 components are distinguished (P3a, P3b), most cases P300 means P3b component. (Figure 6.) The P300 is a positive amplitude EEG wave that occurs during decision-making and information processing. P300 has a peak latency in the range of 250–750 ms. P3b usually appears as a result of a very surprising stimulus, the amplitude of which correlates with the probability, complexity, and form of the appearing stimulus. P300 depends on the energy invested in the task and its complexity. The amplitude of the response increases with the complexity and power of surprise of the task as well.

P3a is sensitive to the context of the stimulus, and its amplitude decreases as the subject becomes accustomed to the stimulus. That is why P3a is not ideal for most BCI applications.

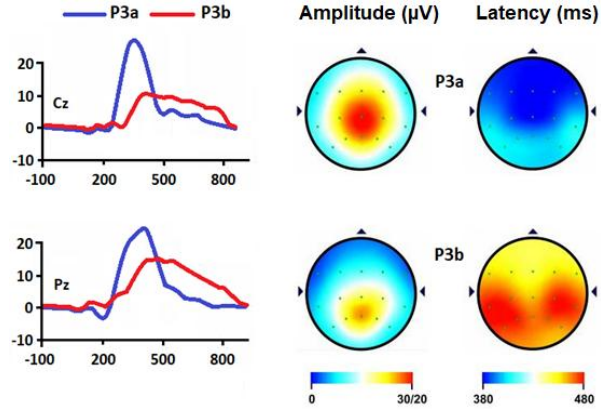


Figure 6.: P3a component is frontocentral and has a peak in the range of 250-280 ms. P3b component is parietal and has peaks around 250-500 ms after the trigger stimulus [14]

The oddball paradigm is used in a typical ERP experiment for evoking P300. (Figure 7.) Where ~80% of the stimuli are standards, and ~20% are deviant. These stimuli can be visual or auditory. Typical tasks are to count the number of deviants. If the presented stimulus is deviant, the subject of the study has to make the decision, and after that, a particular, well-defined task should be made. This decision evokes P300. In BCI applications, multiple stimuli are presented, and investigating the corresponding P300 response in the EEG signal can indicate the person's intentions. P300 spellers usually use this method.

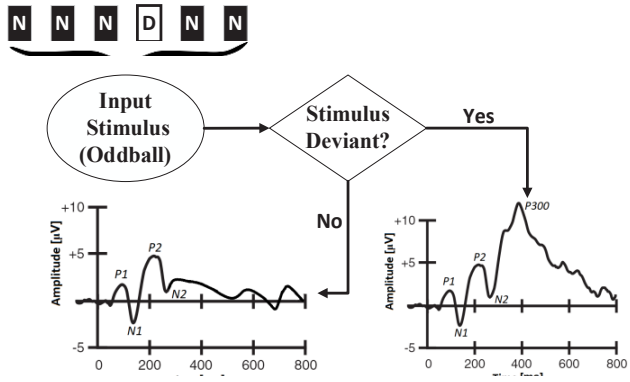


Figure 7.: Oddball paradigm, the subject makes a decision which evokes the P300.

F. Steady-State Visually Evoked Potential

Steady-state visually evoked potential (SSVEP) is a natural response of the visual cortex to a periodic visual stimulus of the retina. The elicited SSVEP response has the same frequency in the EEG signal as the stimulus. SSVEP is strongest at the occipital region (at O1, O2 electrode positions, see Figure 9.). Visual stimulus frequency can be between 3,5–75 Hz.

SSVEP has a high communication rate due to its excellent signal-to-noise ratio, easy configuration, and user training. The limitations of SSVEP in BCI applications are the monitor's (on which the periodic visual stimulus appears) frequency and area. Visual stimulus frequency should be an integer divisor of the monitor frequency.

BCI users are instructed to focus on the corresponding stimulus: a light source or a bright area on a monitor. (Figure 8.) To this stimulus, the visual cortex response (the SSVEP) can be observed in the corresponding electrodes EEG signal.

Only a limited number, around four to six, of stimuli can be efficiently presented to the user. Increasing the number of stimuli creates an overlap in the user's field of vision; thus, the SSVEP generated by the effect of more stimuli appears in the EEG, which reduces the signal-to-noise ratio.

SSVEP has a robust frequency characteristic: the frequency coding method, which assembles different flickering frequencies into multiple targets, has been widely used in BCI applications [15].

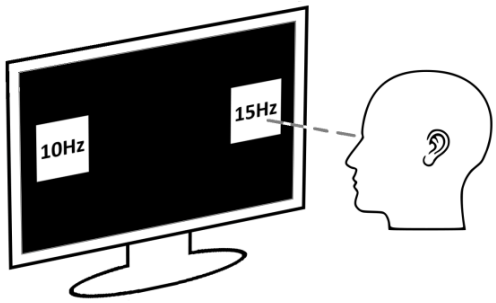


Figure 8.: SSVEP measurement setup, visual stimuli appear on the monitor screen.

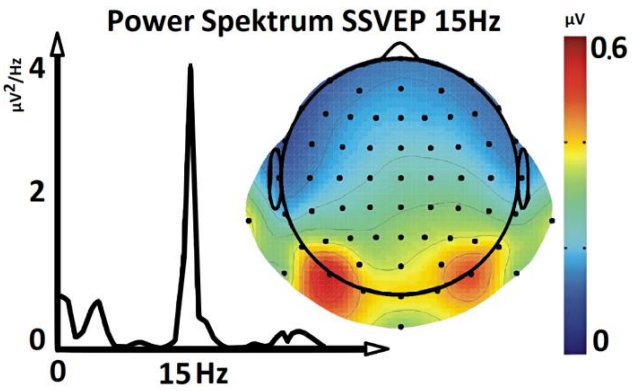


Figure 9.: SSVEP measurement and distribution. [16], [17]

G. Pattern-Reversal VEP

In the clinical use case, visual evoked potentials (VEPs) test the functional integrity of the anterior visual pathways, measured above the visual cortex. For the majority of clinical applications, the pattern-reversal VEP is considered the preferred and most reliable method due to its consistency in timing and waveform, which is less prone to variation compared to other VEP techniques [18]. Therefore, it is ideal for testing purposes.

To evoke Pattern-Reversal VEP event-related potential, the visual cortex is stimulated with a checkerboard visual stimulus (Figure 10.). This visual stimulus alternates a checkerboard image and its inverse in the visual field with a given frequency of 0.5–1.5 Hz. This pattern keeps a constant luminous intensity.

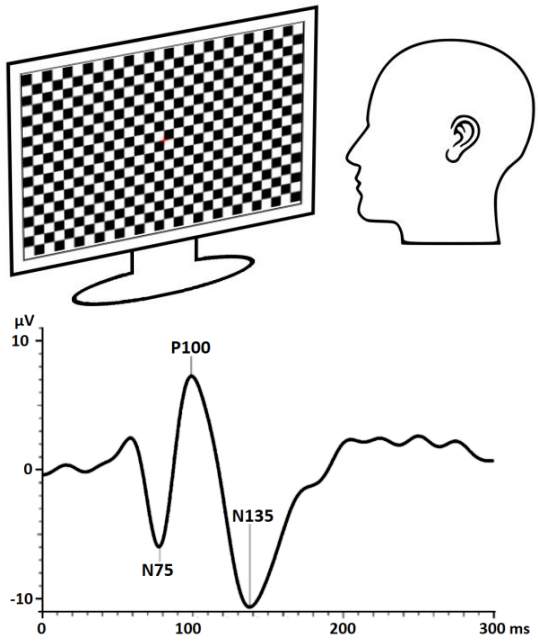


Figure 10.: Chequerboard stimulus and the evoked Pattern-Reversal VEP. [18]

III. RESEARCH MOTIVATION

The primary aim of this research is to develop and validate a robust methodology that can effectively characterize and evaluate different BCI systems, specifically those employing wireless communication interfaces. This validation framework focuses on crucial performance metrics, such as system delay, functional accuracy, and frequency selectivity. Importantly, it also addresses the common challenge of noise artifacts, which can distort EEG measurements and compromise system reliability.

Through systematic validation, this study seeks to demonstrate the applicability of the proposed methodology on a selected BCI system. The validated system will be applied in capturing event-related potentials (ERPs), including steady-state visually evoked potentials (SSVEP), pattern-reversal visually evoked potentials, and the P300 event-related potentials, showcasing its practical utility and reliability in cognitive and clinical EEG assessments.

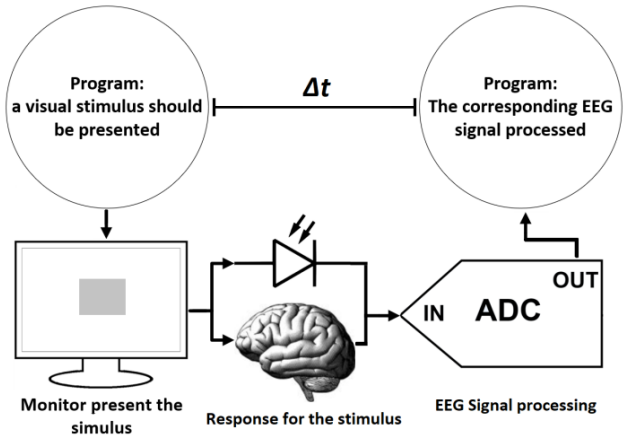


Figure 11.: Validation framework with brain model and in vivo.

IV. VALIDATION METHODOLOGY AND FRAMEWORK

The system architecture of the validation framework can be seen in Figure 11. During the validation and characterization steps, a 24-bit analog-to-digital converter (ADC) was applied, which is directly developed for EEG purposes. The AD converter circuitry consists of a preamplifier, and the digital data is transmitted to the measurement control computer through a wireless fidelity (Wi-Fi) connection. C#-based program running on the computer was responsible for handling and displaying the desired image content (shapes, chequerboard, etc.) and setting the timer to zero. A self-developed MathWorks MATLAB tool was developed and applied to process the incoming signals.

First, the validation framework was tested without electrodes, applying triangle and square waves directly to the inputs. The analog-to-digital converter (ADC) measured the signals' correct morphology, amplitude, and frequency. Further testing of the ADC without the electrodes is not needed at this point. This test shows that the ADC is successfully integrated into the system.

A. Test configuration with the brain model

The brain-computer interfaces that use a photodiode (SFH 2701) operated in photoconductive mode as visual stimuli can also be used as brain models (Figure 12.). The whole system (event generation, visualization, signal measurement, and processing) can be tested with this simple brain model.

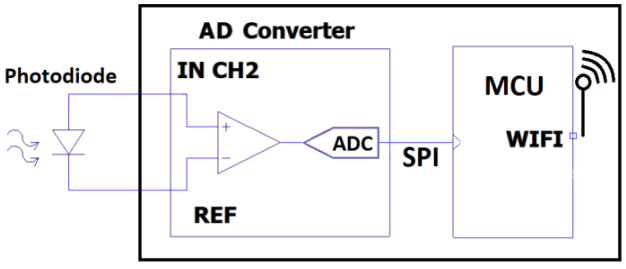


Figure 12.: The test configuration with the brain model.

The brain model creates low-latency and repeatable signals without the noise artifacts created by the brain and the human body. By applying this model, the Δt latency can be measured between the trigger impulse and the measured EEG signal. This method can be used to create a perfect stimulus-ERP synchronization. (Figure 13.)

The brain model measurements show that the tested BCI has a Δt latency with a Gaussian distribution. The latency is random and stays in a 35-55 ms range 94% of the time, as it can be seen in Figure 14. This result is acceptable for most ERP experiments based on the following measurements presented in Section V.

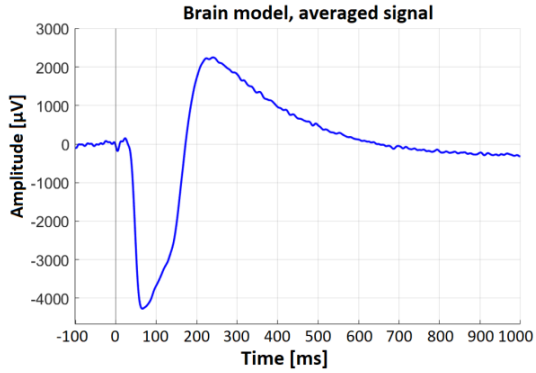


Figure 13.: Brain model measurement to determine the average Δt latency.

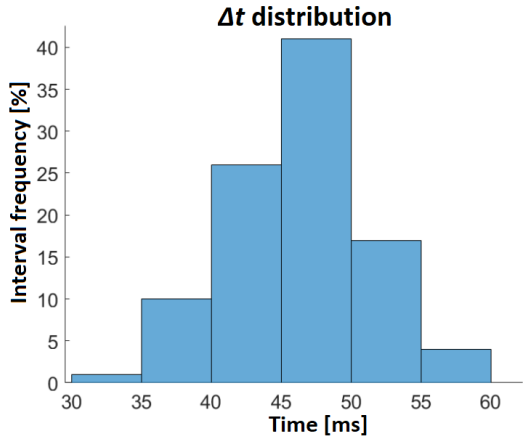


Figure 14.: Brain model measurement Δt latency standard deviation.

However, another acceptance level can be determined based on the morphology of the ERP signal to be measured. More trials lead to significant signal degradation, as it can be seen in Figure 16.

The epochs are defined with trigger impulses. Therefore, the Δt-latency dispersion creates an additional error during the signal averaging (Figure 15 and Figure 16).

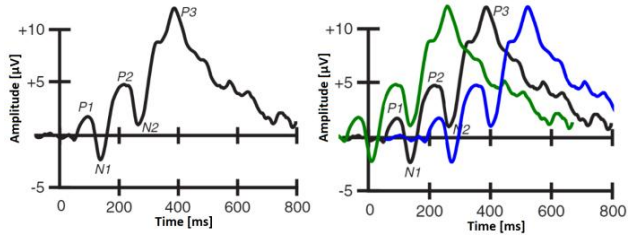


Figure 15.: P300 Δt latency dispersion creates an additional error.

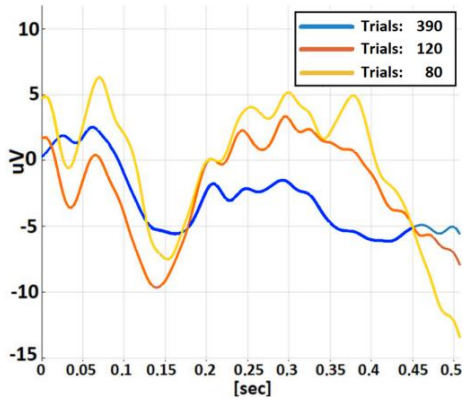


Figure 16.: Signal degradation due to insufficient Δt latency distribution.

B. Test configuration with a BCI device

A Mindrove BCI headset applied in our investigations observes brain activity through six configurable semi-dry platinum electrodes, one reference electrode, and one bias electrode [19], [20]. The lack of conductive gel and the use of platinum instead of silver chloride electrodes lead to an acceptable – but decreased – signal-to-noise ratio, higher electrode impedance and DC offset [21], [22], [23]. On the other hand, omitting the conductive gel and using more durable electrodes creates a better user experience.

The headset applies a 24-bit analog-to-digital converter designed explicitly for EEG measurements with a built-in 24x input gain amplifier and 500 Hz sample frequency. The headset uses 2.4 GHz frequency wireless transmission to connect with the PC client program.

TABLE 1.
EXAMPLE TECHNICAL SPECIFICATIONS REQIREMENTS COMPLIANCE FOR BCI SYSTEMS

	MINIMUM TECHNICAL REQUIREMENTS [1]	Tested Device
Electrodes material	Silver—silver chloride or gold disk electrodes recommended, other materials and electrode pastes can be used.	Platina- Iridium electrode
Electrode placement nomenclature:	10 to 20 System recommended, or 10 to 10 System can be used	10 to 20 System
Electrode Impedace:	100 Ω < R < 10 kΩ with balanced impedances	16 kΩ< R <50 kΩ
Sampling rate	256 Hz minimum, 512 Hz preferable	500 Hz
AD converter resolution	16 bit minimum, 24 bit or more is preferable	24 bit
EEG resolution	<0.05 µV	0.022 µV
Common mode rejection ratio	90 dB minimum, and preferably higher	110 dB
Additional amplifier noise in the recording	< 1µV peak to peak at any frequency 0.5 - 100 Hz, including at 60 Hz	1.39 µV peak to peak
Trigger Latency	not specified	35-55 ms

V. VALIDATION OF WIRELESS BCI WITH EVENT-RELATED POTENTIALS

Open/closed eye alpha rhythm (Berger effect) measuring is usually the widely used way – and the first step – to test basic functionality of the validation framework. Alpha wave is in a relatively narrow frequency range. However, amplitude can be varied by the mental process and mental states [3] SSVEP amplitude can also be affected by these, but it is considered acceptable for BCI device validation.

A. SSVEP

Steady-state visually evoked potential (SSVEP) is a robust, noise-tolerant ERP, measured over the visual cortex at O1 and O2 electrode positions. SSVEP is a time-independent ERP. The evoked response can be easily distinguished in the EEG amplitude spectrum. The ERP processing algorithm is based on computing the discrete Fourier transform. Therefore, SSVEP is optimal for the BCI test, where periodic noises could occur in the EEG signal. A lower frequency (<20Hz) stimulus creates a higher amplitude response. SSVEP is most sensitive to 15 Hz stimulus frequency. BCI was tested with a flashing LED light and a rectangle on the monitor, and measured at the subject's O1 electrode position. The tested BCI measured SSVEP correctly in both experiments. The EPR signals are present in the EEG recording with adequate amplitude (Figure 17. and Figure 18.).

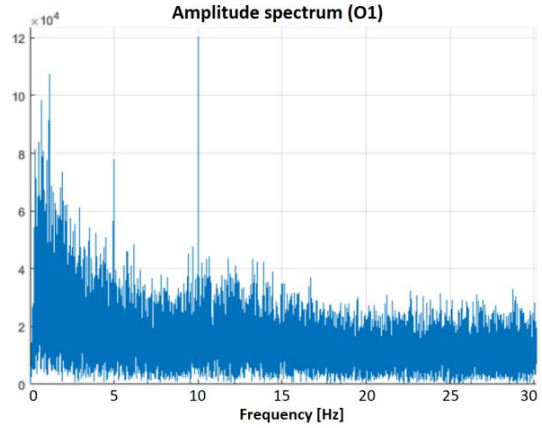


Figure 17.: SSVEP measurement with 15 Hz LED.

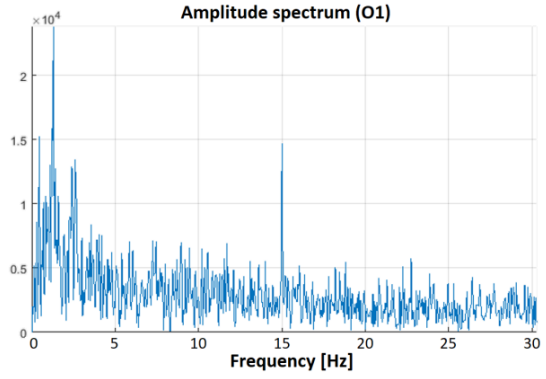


Figure 18.: SSVEP measurement with 10 Hz white rectangle.

B. *Pattern-Reversal VEP*

Pattern-Reversal VEP experiment focuses on the correct timing. And also the recommended testing ERP for clinical use EEG systems, by the American Clinical Neurophysiology Society [1]. The ERP signal had a definite ~30 ms wide positive amplitude. Pattern-Reversal VEP was also measured at O1, and the electrode placement was the same as SSVEP discussed before. The signal-to-noise ratio can be easily increased by incrementing the number of epochs. 1 Hz stimulus frequency creates plenty of ERP for measurements.

This ERP measurement shows whether the trigger impulse and the observed event timing are synchronized.

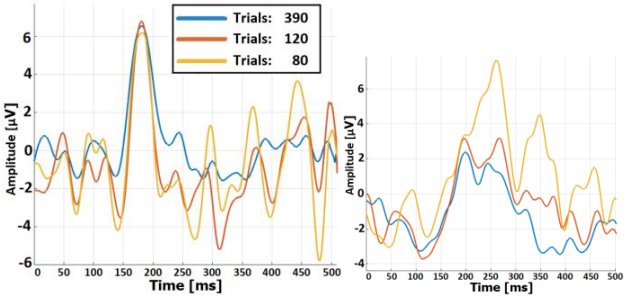


Figure 19.: Pattern-Reversal VEP measurements
Averaging different numbers of trials.
(left: small standard deviation Δt, right: large standard deviation Δt)

In Figure 19., the left figure shows the Pattern-Reversal VEP with correct timing, and the right one shows another Pattern-Reversal VEP measurement, where Δt latency stays in a 50 ms range 50% of the time. The inaccuracy of Δt stretches out the detected ERP signals, and it can be seen in Figure 19. that the maximums of different numbers of trials are at different places on the function.

C. *P300 ERP measurements*

P300-based spellers are widely used due to the relatively high information throughput. P300 is evoked via a visual oddball paradigm shown in Figure 20. P300 is sensitive to the stimulus, and a familiar face can evoke a more distinct signal modality and better evaluation speed [24].

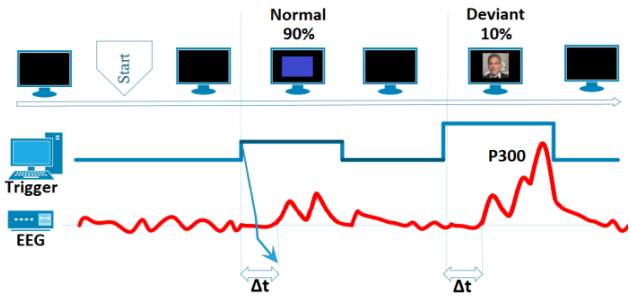


Figure 20.: P300 oddball paradigms stimuli.

The oddball paradigms stimuli were presented with 1 Hz frequency, and stimuli were presented at the monitor for 0.1 sec. The probability of the deviant paradigm was Pd=0.1. (Figure 20.)

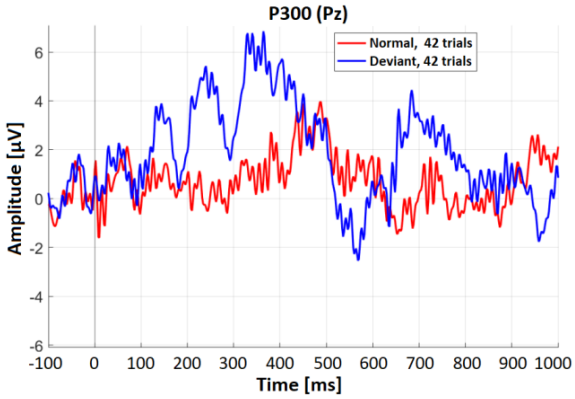


Figure 21.: P300 measurement platina electrode + NaCl.

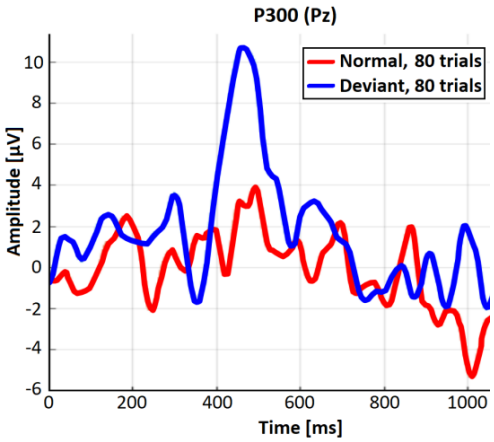


Figure 22.: P300 measurement platina electrode + Abralyt HiCl 10% gel

The measurements were recorded at a sampling rate of 500 Hz and band-pass filtered between 1 and 90 Hz. Power line interference at 50 Hz was eliminated, and additional artifacts were manually removed from the EEG signal. (Figure 21. and Figure 22.)

In Figure 22., it can be seen clearly that the deviant stimuli caused higher, detectable amplitude at around 300 ms, and the amplitude is approximately three times greater than the amplitude of the normal stimuli curve.

VI. CONCLUSION

The primary aim of our work was to develop a methodology strictly focusing on ERP examinations. Thus, it is possible to determine whether the BCI system under investigation can be used in ERP experiments.

- The ACNS Guidelines outline the minimum system requirements for clinical EEG systems. While these standards should be applied to BCI systems, their use is not mandatory.
- Crucial performance metrics such as system delay, functional accuracy, and frequency selectivity should be validated for the whole system with the corresponding ERPs.

- SSVEP for frequency selectivity validation. It is one of the mainly used ERP in BCI systems, and is not affected by system's overall latency.
- Pattern reversal VEP due to low variability in timing, amplitude, waveform morphology, and validation can be performed at relatively high stimulus frequency.
- The presented brain model can be used to determine the system's overall latency (Δt), and its distribution.
- P300 is the most common ERP used by BCI systems, amplitude, waveform morphology can be determined. However, it exhibits greater variability than Pattern reversal VEP.

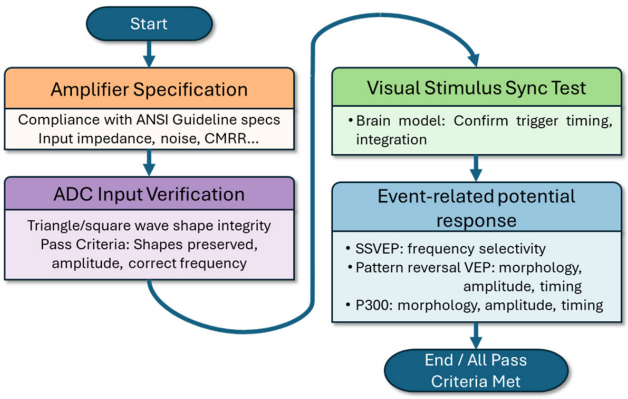


Figure 23.: Measurement and validation steps.

Measurement and validation steps were carried out, and the results proved the applicability of the validation methodology. The developed validation methodology can be seen in Figure 23. It includes not only the mandatory steps but also the key characteristic parameters that must comply with the technical specification requirements for BCI systems. The characterized BCI system was proved to be applied in ERP EEG measurements such as steady-state visually evoked potential, pattern-reversal visually evoked potential, and P300 event-related potential. The developed validation methodology is suitable for testing and validating similar BCI systems.

ACKNOWLEDGMENT

The research presented in this paper was fully supported by the K_20 grant of the National Research, Development and Innovation Office (NKFIH) through the project No. 135224.

REFERENCES

[1] American Clinical Neurophysiology Society, "ACNS Guideline," American Clinical Neurophysiology Society Guideline. Accessed: Jun. 01, 2025. [Online]. Available: https://www.acns.org/practice/guidelines

[2] A. Craik *et al.*, "Design and Validation of a Low-Cost Mobile EEG-Based Brain–Computer Interface," *Sensors*, vol. 23, no. 13, p. 5930, Jun. 2023, DOI: 10.3390/s23135930.

[3] M. Lopez-Gordo, D. Sanchez-Morillo, and F. Valle, "Dry EEG Electrodes," *Sensors*, vol. 14, no. 7, pp. 12 847–12 870, Jul. 2014, DOI: 10.3390/s140712847.

[4] K. Jakab, J. Csipor, I. Ulbert, Z. Keresztes, G. Mészáros, and G. Márton, "EEG sensor system development consisting of solid polyvinyl alcohol–glycerol–NaCl contact gel and 3D-printed, silver-coated polylactic acid electrode for potential brain–computer interface use," *Materials Today Chemistry*, vol. 26, p. 101 085, Dec. 2022, DOI: 10.1016/j.mtchem.2022.101085.

[5] W. Fadel, M. Wahdow, C. Kollod, G. Marton, and I. Ulbert, "Chessboard EEG Images Classification for BCI Systems Using Deep Neural Network," in *Bio-inspired Information and Communication Technologies*, vol. 329, Y. Chen, T. Nakano, L. Lin, M. U. Mahfuz, and W. Guo, Eds., in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 329., Cham: Springer International Publishing, 2020, pp. 97–104. DOI: 10.1007/978-3-030-57115-3_8.

[6] S. J. Luck, "Applied Event-Related Potential Data Analysis," *2022, LibreTexts*. DOI: 10.18115/D5QG92.

[7] M. J. Giuliodori and G. Zuccolilli, "POSTSYNAPTIC POTENTIAL SUMMATION AND ACTION POTENTIAL INITIATION: FUNCTION FOLLOWING FORM," Advances in *Physiology Education*, vol. 28, no. 2, pp. 79–80, Jun. 2004, DOI: 10.1152/advan.00051.2003.

[8] R. B. Bod, J. Rokai, D. Meszéna, R. Fiáth, I. Ulbert, and G. Márton, "From End to End: Gaining, Sorting, and Employing High-Density Neural Single Unit Recordings," *Front. Neuroinform.*, vol. 16, p. 851 024, Jun. 2022, DOI: 10.3389/fninf.2022.851024.

[9] D. Purves and S. M. Williams, Eds., *Neuroscience*, 2nd ed. Sunderland, Mass: Sinauer Associates, 2001.

[10] J. N. Acharya, A. Hani, J. Cheek, P. Thirumala, and T. N. Tsuchida, "American Clinical Neurophysiology Society Guideline 2: Guidelines for Standard Electrode Position Nomenclature," *Journal of Clinical Neurophysiology*, vol. 33, no. 4, pp. 308–311, Aug. 2016, DOI: 10.1097/WNP.0000000000000316.

[11] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, "The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology," *Electroencephalogr Clin Neurophysiol Suppl*, vol. 52, pp. 3–6, 1999.

[12] S. Sur and V. Sinha, "Event-related potential: An overview," *Ind Psychiatry J*, vol. 18, no. 1, p. 70, 2009, DOI: 10.4103/0972-6748.57865.

[13] S. J. Luck, An introduction to the event-related potential technique, Second edition. Cambridge, Massachusetts: The MIT Press, 2014.

[14] M. A. Conroy and J. Polich, "Normative Variation of P3a and P3b from a Large Sample: Gender, Topography, and Response Time," *Journal of Psychophysiology*, vol. 21, no. 1, pp. 22–32, Jan. 2007, DOI: 10.1027/0269-8803.21.1.22.

[15] X. Chen, Y. Wang, S. Zhang, S. Xu, and X. Gao, "Effects of stimulation frequency and stimulation waveform on steady-state visual evoked potentials using a computer monitor," *J. Neural Eng.*, vol. 16, no. 6, p. 066 007, Oct. 2019, DOI: 10.1088/1741-2552/ab2b7d.

[16] M. J. Wieser, L. M. McTeague, and A. Keil, "Sustained Preferential Processing of Social Threat Cues: Bias without Competition?," *Journal of Cognitive Neuroscience*, vol. 23, no. 8, pp. 1973–1986, Aug. 2011, DOI: 10.1162/jocn.2010.21566.

[17] K. Saupe, "Neural mechanisms of intermodal sustained selective attention with concurrently presented auditory and visual stimuli," *Front. Hum. Neurosci.*, vol. 3, 2009, DOI: 10.3389/neuro.09.058.2009.

[18] R. S. Al-Najjar, N. M. Al-Aubody, S. Z. Al-Asadi, and M. Alabbood, "Pattern-Reversal Visual Evoked Potentials Tests in Persons with Type 2 Diabetes Mellitus with and without Diabetic Retinopathy," *Neurology Research International*, vol. 2020, pp. 1–7, Aug. 2020, DOI: 10.1155/2020/1014857.

[19] "Mindrove ARC." Accessed: Jun. 03, 2024. [Online]. Available: https://mindrove.com/arc/

[20] E. Noboa, M. Rácz, L. Szűcs, P. Galambos, G. Márton, and G. Eigner, "Development of an EMG based SVM supported control solution for the PlatypOUs education mobile robot using MindRove headset," *IFAC-PapersOnLine*, vol. 54, no. 15, pp. 304–309, 2021, DOI: 10.1016/j.ifacol.2021.10.273.

[21] P. Tallgren, S. Vanhatalo, K. Kaila, and J. Voipio, "Evaluation of commercially available electrodes and gels for recording of slow EEG potentials," *Clinical Neurophysiology*, vol. 116, no. 4, pp. 799–806, Apr. 2005, DOI: 10.1016/j.clinph.2004.10.001.

[22] C. Guger *et al*., "How many people are able to control a P300-based brain–computer interface (BCI)?," *Neuroscience Letters*, vol. 462, no. 1, pp. 94–98, Sep. 2009, DOI: 10.1016/j.neulet.2009.06.045.

[23] C. Guger, G. Krausz, B. Z. Allison, and G. Edlinger, "Comparison of Dry and Gel Based Electrodes for P300 Brain–Computer Interfaces," *Front. Neurosci*., vol. 6, 2012, DOI: 10.3389/fnins.2012.00060.

[24] C. Guger, R. Ortner, S. Dimov, and B. Allison, "A comparison of face speller approaches for P300 BCIs," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, Hungary: IEEE, Oct. 2016, pp. 004 809–004 812. DOI: 10.1109/SMC.2016.7844989.

Ádám Salamon received his M.Sc. degree in Electronic Engineering from the Budapest University of Technology and Economy (BME), Faculty of Electrical Engineering and Computer Science, Hungary, in 2022. He won the 1st prize in Students' Scientific Conference at the Faculty of Electrical Engineering and Computer Science at BME in 2020. He started his PhD study at the Department of Electron Devices, BME in 2022. His actual research field is brain-computer interfaces and EEG sensor systems, which is based on brain waves, event-related potentials, and emotion detection. Also currently working on new types of EEG electrodes.

Gábor Takács obtained his M.Sc. and Ph.D. degrees in Electrical Engineering at the Budapest University of Technology and Economics in 2013 and 2018, respectively. He started working at the Department of Electron Devices, BME, in 2015 as an assistant lecturer and was promoted to associate professor in 2024. Since 2013, he has been dealing with the characterization of hydrodynamic and thermal performance of mini- and microscale cooling structures. Another interest is the analog CMOS integrated circuit design. He is a professor in the Joint International Master in Smart Systems Integrated Solution Erasmus Mundus program. Since 2019, he has been a member of the Micro- and Nanoelectronics Section of the Scientific Association for Infocommunications. As of 2020, he has been a work package leader in a national fundamental research project No. 135224 (Novel solutions for thermal issues in chiplet-based System-on-Package devices) funded by the National Research, Development, and Innovation Office (NKFIH). He has been working as a technical expert in the European Union's Horizon 2020 funded project DigiFed No. 872088.

György Bognár received PhD and M.Sc. degrees in Electronic Engineering from Budapest University of Technology and Economy (BME), Faculty of Electrical Engineering and Computer Science, Hungary, in 2010 and 2002, respectively, and M.Edu. degree in Teacher of Electrical Engineering from BME, Faculty of Economic and Social Sciences, Hungary, in 2008. He earned M.A. in Talent Development at Károly Eszterházy University, Eger, Hungary, in 2010. He joined the Department of Electron Devices as an Assistant Lecturer in 2004. Since 2010, he has been an Associate professor at the Department of Electron Devices, BME, Hungary. Since 2019, he has been the head of the Micro- and Nanoelectronics Section of the Scientific Association for Infocommunications. His actual research field is thermal management of System-in-Package devices, thermal simulation, modeling, and technology of heterogeneously integrated 2.5D and 3D packages with embedded thermal management solutions. Since 2007, he has been dealing with the realization technology, hydrodynamic and thermal modeling of microscale cooling structures, and microchannel heatsinks.
Between 2014 and 2018, Dr. Bognár was the supervisor of a national fundamental research project (Integrated thermal management in System-on-Package devices) funded by the Hungarian Scientific Research Fund (OTKA). Since 2020, he has been the supervisor of a national basic research project (Novel solutions for thermal issues in chiplet-based System-on-Package devices) funded by the National Research, Development, and Innovation Office (NKFIH). He worked on several national and international projects as a researcher.

Masked Face Image Inpainting Based on Generative Adversarial Network

Qingyu Liu, Lei Chen, Yeguo Sun*, and Lei Liu

***Abstract*—Face image inpainting is a critical task in computer vision due to the intricate semantic and textural features of facial structures. While existing deep learning-based methods have achieved some progress, they often produce blurred or artifact- prone results when handling large occlusions, such as face masks. To address these challenges, this paper proposes a novel generative adversarial network (GAN) framework tailored for masked face inpainting. The generator adopts a U-Net architecture enhanced with a multi-scale mixed-attention residual module (MMRM), which integrates multi-branch convolutions for diverse receptive fields and combines spatial-channel attention mechanisms to prioritize semantically relevant features. The decoder further enhances feature fusion through channel attention mechanism, which selectively emphasizes meaningful patterns during feature map reconstruction. A realistic masked face dataset is synthesized using the CelebA database by dynamically adjusting mask positions, sizes, and angles based on facial landmarks, ensuring alignment with real-world scenarios. Quantitative and qualitative evaluations demonstrate that our method outperforms conventional models in both visual quality and quantitative metrics. Ablation studies further validate the effectiveness of MMRM and attention mechanisms in preserving structural coherence and reducing artifacts.**

***Index Terms*—Attention mechanism, Generative adversarial network, image inpainting, residual module**

I. INTRODUCTION

Face recognition, a type of biometric technology, has been widely used in different areas of social life, and some excellent face recognition technologies have even surpassed the level of human recognition. However, the current face recognition accuracy is still affected by other external factors, such as mask blocking. When the face is obscured by a mask, the effective features that can be extracted from the face are greatly reduced, which leads to a lower accuracy rate of face recognition. Image inpainting is the process of inferring and reconstructing the content of a missing area from limited known image information, making the image visually complete and natural. Using image inpainting techniques to remove face occlusion and try to restore the face to its initial state has become a hot issue of concern. This research presents a new generative adversarial network (GAN) model for face inpainting.

Image inpainting techniques, in terms of their development process, have two main categories: traditional techniques and deep learning-based techniques. The traditional image inpainting techniques are mainly diffusion-based [1], [2] and block-based methods [3], [4]. The primary principle of the diffusion inpainting algorithms is to disperse the pixels from the image's unobstructed regions to the obstructed regions to achieve the goal of inpainting. Bertalmio et al. [5] were the first to apply the method to image inpainting by proposing the BSCB model, which used the heat diffusion equation in fluid dynamics to propagate domain information into the occluded regions. Efros and Leung et al. [6] proposed a Markov random field-based approach to fill the occluded regions by selecting the best matching sample blocks from the unoccluded regions. In order to reduce the cost of searching and matching, the PatchMatch approach was presented by Barnes et al. [7] and used a quick nearest neighbor algorithm for searching. Such methods have an assumption that the sample image contains similar information about the occluded regions. This is not guaranteed in many cases. In addition, the block-based inpainting methods rely on a large training dataset and are too computationally intensive.

In 1998, the LeNet model, one of the first convolutional neural network (CNN) models, was introduced by LeCun et al. [8]. This model catalyzed the advancement of CNN. Image inpainting techniques based on deep learning have been rapidly developed since the advent of CNN. The methods of deep learning [9], [10], [11], [12] can effectively extract the features of the images and fill the occluded areas of the images, to achieve the purpose of inpainting. In 2014, Goodfellow et al. [13] pioneered GAN. GAN pushed image processing to a more advanced stage of development. In the model, the discriminator and generator are trained against one another, and the two improve their respective abilities during the training, as shown in Figure 1. Using the trained generator we can fit real data distributions to achieve tasks such as image generation and image inpainting. In the year 2016, Pathak et al. [14] first used GAN for image inpainting by proposing the Context Encoder model. This model employs an encoder-decoder network topology, in which the encoder is used for image feature extraction and the decoder creates an image the

The work is supported in part by the Key Projects of Natural Science Research in Anhui Colleges and Universities under Grant 2023AH051546, in part by the University Natural Science Foundation of Anhui Province under Grant 2022AH010085, in part by the Opening Foundation of State Key Laboratory of Cognitive Intelligence under Grant COGOS-2023HE02, in part by Program of Anhui Education Department under Grant 2024jsqygz83.

Qingyu Liu, Lei Chen and Lei Liu are with Faculty of School of Computer Science, Huainan Normal University, Huainan, China (e-mail: liuqy@hnnu.edu.cn, LeiChn0911@163.com, 357920185@qq.com).

*Yeguo Sun is with Faculty of School of Finance and Mathematics, Huainan Normal University, Huainan, China (*Corresponding Author, e-mail: yeguosun@126.com).

DOI: 10.36244/ICJ.2025.2.2

same size as the obscured image. Regularly obscured images can be restored using the Context Encoder model, although the restored images have issues including blurring and inconsistent content. Subsequently, Iizuka et al. [15] proposed a bi-discriminators GAN model. The generated images and real images are fed into the global discriminator and the images are evaluated as a whole for consistency to ensure the overall coherence of the images. The use of local and global discriminators enables the generator to better learn the data distribution of real images, thereby improving the model's restoration capabilities. Yu et al. [16] introduced a new end-to-end GAN for image inpainting using coarse and fine inpainting networks in tandem. The coarse inpainting stage accomplishes the initial contour generation of the occluded image, and the fine restoration realizes the image fine processing on this basis, which ultimately results in a high-quality inpainting result.

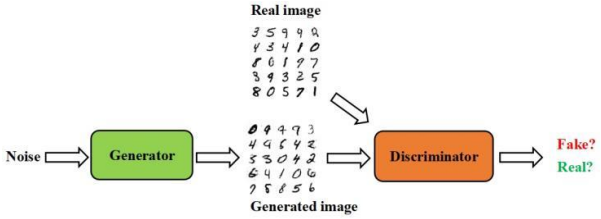


Fig. 1. GAN structure.

Face image inpainting is a sub-task of image inpainting, but it is special due to the complex texture structure and rich semantic features of the face [17], [18], [19]. GAN were initially used in the field of face image inpainting by Li et al. [20]. He proposed the concept of semantic parsing loss in the loss functions and adopted a pre-trained parsing network for face image inpainting. Nazeri et al. [21] introduced a 2-step inpainting method in which the first step predicts and restores the contours of the face and then restores the detailed aspects of the face. Qin et al. [22] proposed a prescription based on weighted face similarity for restoring face images with relatively large occluded regions. Yang et al. [23] used a bi-discriminator to restore occluded face images to maintain better semantic consistency. Although these improved face image inpainting methods based on GAN can perform face image inpainting, there are still some problems. For example, how to ensure the balanced training of the generator and discriminator, how to realize the fine inpainting of face images and reduce the complexity of the model, how to avoid using extra input information such as edge contours in the face inpainting process, and so on. To address these issues, we proposed a GAN model based on multi-scale mixed-attentions residual module (MMRM). The following are the primary contributions of this work:

(a) Based on the public CelebA dataset, the masks are added by accurately detecting the positions of the face feature points. The position, size and angle of the face in each image are different, and thus the position, size and angle of the face mask will also change. Therefore, the produced dataset is more in line with the actual situation of people wearing masks, and the trained generator model has stronger generalization ability.

(b) The generator model is based on the U-Net structure [24], and the feature fusion of down-sampled texture features and up-sampled high-level semantic features can be performed through skip-layer connections to improve the image inpainting of faces.

(c) We use different scales of convolutional branches for the residual module. Each branch uses spatial attention [25] to extract features, and multiple branch features are fused before adding channel attention [26] to focus on more effective channel features.

II. Proposed Method

GAN was originally proposed in 2014 and the original inputs to the generator are random vectors. We use GAN for face image inpainting, so the inputs to the generator are the face images wearing masks and the outputs are the restored face images without masks. Figure 2 displays the model's general architecture. The generator in the model makes use of the U-Net structure, and the discriminator is also part of the model. U-Net has the capacity to extract features at various scales in the encoding phase and fuse the extracted features in the decoding phase. This network can extract face semantic information quickly and makes training easier and convergence faster. The discriminator adopts the PatchGAN [27] structure and uses a spectrally normalized convolutional layer instead of the normal convolutional layer, which improves the texture of the restored region and has the effect of stabilizing the training. The generator uses MMRM for both the modules of up-samplings and down-samplings.

A. Generator

Because of its skip connections, the U-Net structure was initially used in the field of medical image segmentation, preserving both high-level semantic information and low-level characteristics. Thus, we combine the U-Net structure with GAG. We improve the original residual module by setting it to three convolutional branches, each of which uses a convolutional kernel of different sizes to extract features separately. Meanwhile, each branch uses a spatial attention module to enhance feature extraction. Figure 3 shows the structure of the improved residual module.

Parameter updates in CNN rely on backpropagation of the gradient. As the depth of the network gradually increases, parameters less than 1 are multiplied together resulting in gradient vanishing. When the network weights are initialized to a larger value, the gradient grows exponentially as the neural network deepens, triggering gradient explosion. Although the gradient problem can be solved to a certain extent by introducing activation functions such as ReLU and Batch Normalization layer, when the depth of the network is increased, it still occurs that the training effect of the deeper network is rather worse than that of the shallower network, i.e., the network degradation problem. He et al. [28] proposed ResNet network in 2016 for solving the degradation problem that occurs when the network depth increases. Thus, both the up-sampling modules and down-sampling modules of the generator use improved residual modules.

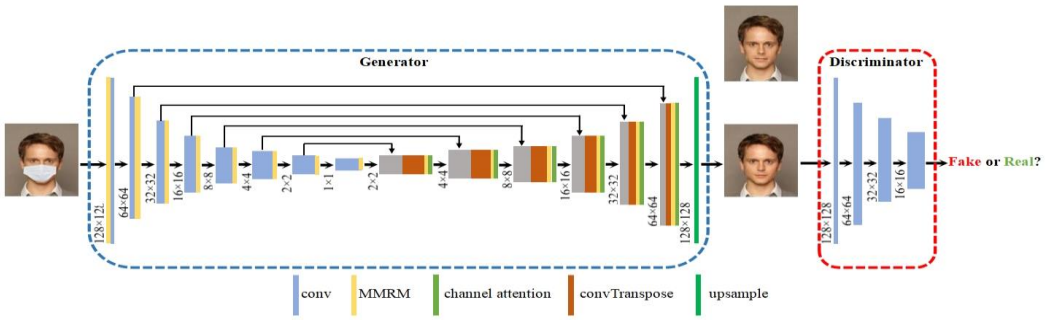


Fig. 2. Architecture of the model.

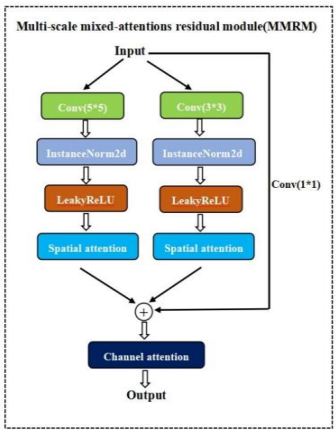


Fig. 3. Architecture of MMRM.

The improved residual module comprises three parallel convolutional branches. The convolution kernel sizes of the three convolution branches are 1×1, 3×3, 5×5, and they have different receptive fields so that features of different scales can be extracted. Spatial attention is introduced in 3×3 and 5×5 branches to dynamically weight the spatial location of the feature map. For example, when restoring mask-obscured regions, the model prioritizes attention to unobscured semantic key points (e.g., bridge of the nose) to enhance the guidance of contextual information. After multi-branch feature splicing, important channel features are filtered by channel attention. For example, channels corresponding to high-frequency information such as facial skin color and hair texture are enhanced, while redundant background channels are suppressed to improve the semantic consistency of the inpainting.

Figure 3 also shows that the spatial attention layers are added to the 3×3 and 5×5 convolutional branches, respectively. The goal of face inpainting is to fill the occluded region with information from the non-occluded regions. Considering that the face image has obvious geometric features, it is obvious that different regions of the face have different effects on the final inpainting result. The final output of the residual module is directly related to the input. Additionally, it is necessary to consider whether the weights of each convolutional branch should be identical. Thus, three convolutional branches use channel attention after the channel dimension is concatenated, focusing on important information to suppress irrelevant information.

B. *Discriminator*

In the image inpainting task, the most important thing is the prediction of the pixel values of the masked region. In adversarial learning, a discriminator constrains the training of the generator by determining whether the images are similar to the real. A good discriminator maintains the overall semantic consistency of the restored images and enhances the texture details of the restored regions. In Figure 4, the network structure is depicted and we employ PatchGAN as the discriminator structure. The discriminator network consists of a stack of 4 convolutional layers with convolutional kernel size 3*3 and stride size 2. The purpose is to obtain the feature statistics of Markov Patch (MP). Each point in the output matrix represents a region of the original input image, by which the efficiency of the discriminator can be improved.

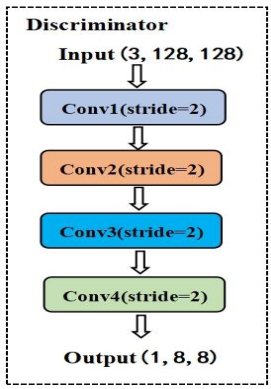


Fig. 4. Discriminator.

C. *Loss functions*

The discriminator and generator are trained adversarially using loss functions. The generator and discriminator iteratively improve their respective capabilities until reaching a steady state. Therefore, loss functions are crucial for GAN. To make the generator better at fixing occluded face images, the generator loss functions use adversarial loss, L1 loss, TV Loss, and they are defined as below:

$$Loss_{adv} = -E_{z\sim P_z} D(G(z)) \quad (1)$$

$$Loss_{L1} = E_{z\sim P_z,\ x\sim P_{data}} \| G(z) - x \| \quad (2)$$

where z denotes the masked face image and x denotes the real face image. D and G represent the generator and discriminator respectively. Considering the coherence of the restored images, we also used the TV loss:

$$Loss_{TV} = \sum_{i,j} \sqrt{(I_{i,j-1} - I_{i,j})^2 + (I_{i+1,j} - I_{i,j})^2} \quad (3)$$

where $I_{i,j}$ denotes a pixel point of the image, and $I_{i,j-1}$, $I_{i+1,j}$ are the neighboring pixel points in the vertical and horizontal directions, respectively. Thus, the generator's loss function is expressed as follows:

$$Loss(G) = \gamma_{adv} * Loss_{adv} + \gamma_{L1} * Loss_{L1} + \gamma_{TV} * Loss_{TV} \quad (4)$$

Where $Loss_{adv}$ represents the adversarial loss function, $Loss_{L1}$ represents the L1 loss function, and $Loss_{TV}$ represents the TV Loss function. The weights of the different loss functions are shown by the symbols γ_{adv}, γ_{L1}, and γ_{TV}, respectively.

The discriminator then has only the adversarial loss function, as follows:

$$\text{Loss(D)} = Loss_{adv} \quad (5)$$

III. Experiments and Results

In this section, we produced a dataset of faces wearing masks. Then, quantitative and qualitative experiments are conducted to compare with other end-to-end image inpainting models such as Pix2Pix [29]. Finally, to show how effective the improved model is, ablation experiments are carried out.

A. Dataset

An excellent dataset for face processing is the CelebA image dataset, an open dataset created by the Chinese University of Hong Kong that includes 202599 images of over 10000 individuals from all over the world. In this study, 40000 of these images were chosen at random to make up the initial dataset, and the face recognition library "Face_recognition" is used to accurately locate the face feature points to complete the addition of masks. Considering the hardware limitations, all images are scaled to 128×128 size uniformly.

The "Face_recognition" library could detect faces in an image and obtain the locations of the key points such as eyes, nose, mouth, etc., as shown in Figure 5. To realistically produce a dataset that matches the real situation, the mask should be able to cover the mouth, chin, and tip of the nose of the face. Also, considering the different positions and angles of the face in the image, the angle and size of the face mask should be changed accordingly. This is extremely important. The size, position and angle of each face may vary. We can adaptively add masks based on the key points of the specific face image. Thus, the produced dataset has better diversity, and the trained model has better adaptability. The specific process of adding a mask to a face is shown in Algorithm 1.

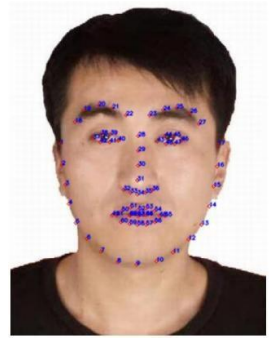


Fig. 5. Key points of the face. There are a total of 68 feature points, each of which is numerically numbered, and the coordinates of each feature point can be obtained.

Algorithm 1 Steps to Produce Masked Face Images

Input: face image without mask
Output: masked face image
1 Detect faces and get all the key points' coordinates;
2 Calculate the distance between point 29 of the nose and point 8 of the chin, and then multiply by 1.2 to get the height of the new mask, denoted as H;
3 Calculate the distance between point 29 of the nose and point 1 of the right side of the face, and then multiply by 1.2 to get the width of the right half of the new mask, denoted as W_R;
4 Intercept the right half of the original mask and transform the size according to (H, W_R);
5 Calculate the distance between point 29 of the nose and point 16 on the left side of the face and multiply by 1.2 to get the width of the left half of the new mask, denoted W_L;
6 Intercept the left half of the original mask and transform to (H, W_R);
7 Merge the left and right sides of the new mask to get a completely new mask;
8 Place the new mask using the center node between point 29 of the nose and point 8 of the chin as a reference point;
9 Determine if nose point 33 and chin point 8 are perpendicular, if not, do the same angular transformation using the geometric center point of the new mask as a datum.

In the above process, the height and width of the new face mask are multiplied by 1.2 from the original calculated values. This is because the mask is meant to cover the key locations such as the nose and mouth, and the values should be slightly larger than the calculated distance. This produces a more realistic image. Figure 6 shows the complete process of adding the mask, following the same steps as above. The input to the algorithm is a normal face, followed by the generation of the right half of the mask and the left half of the mask. Then later on it is spliced to form a complete mask and covered in the right place of the face. The last step is to adjust the angle of the mask. Following this method, we choose the three most common masks in daily life to produce the dataset images, as shown in Figure 7.

B. Experimental environment and parameter settings

The experimental hardware environment is an Intel Xeon processor, with 128GB RAM and 8 NVIDIA RTX 3090 Ti GPU cards. The software environment uses Red Hat 4.8.5, Python 3.7.1, and PyTorch 1.10. The model is trained using data parallelism, in which the dataset images are evenly distributed across all computational cards. During the training process, each node updates its model parameters independently, and

finally, the model parameters of each node are fused to obtain the final model. In this way, the computation and memory consumption of a single node can be greatly reduced, and the overall training speed of the model can be improved.

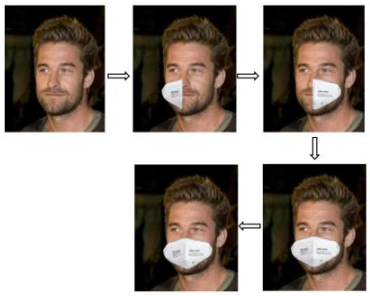


Fig. 6. Mask addition process.

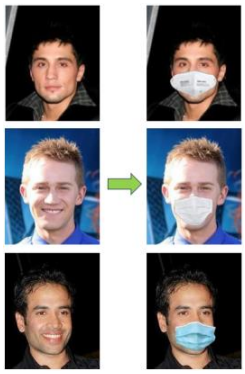


Fig. 7. Masked images with three types of face masks.

The experimental hyperparameters are set in Table I. The weights of the generator adversarial loss, L1 loss, and TV loss are 0.01, 0.95, and 0.04, respectively. The model optimizer uses Adam with an initial learning rate of 0.0002, and the batch size is set to 1500 according to the GPU memory capacity. For later comparison experiments, all the image inpainting models are trained for a maximum of 1000 epochs.

The model took 19 hours for a single training process, and used about 23GB of each GPU's memory (for a total of 24GB on a single card).

TABLE I
TRAINING HYPERPARAMETERS

Parameter	Value	Parameter	Value
γ_{adv}	0.01	Learning rate	0.0002
γ_{L1}	0.95	Epochs	1000
γ_{TV}	0.04	Optimizer	Adam
Batch size	1500		

C. Experimental results

In this section, through some contrast experiments, we qualitatively and quantitatively evaluate the results of our proposed method. The models we use as comparisons are Pix2Pix [30], Shift-Net [31], GMCNN [32], WaveFill [33], etc.

1) Qualitative analysis

Figure 8 displays the comparison results using our new dataset. The first columns are the initial face images without masks and the second columns are the images after the masks are automatically added according to Algorithm 1. The next five columns show the comparison results of different inpainting methods. Pix2Pix uses the U-Net structure to reconstruct face images, but it doesn't work well for face images with complex image structures. So we can see obvious blurred pixels. Shift-Net adds a shift-connection layer to the generator network based on the Pix2Pix model and adds guidance loss to the overall loss functions, so it gives better inpainting results than Pix2Pix. The restored regions (e.g., nose and mouth) exhibit significantly better quality compared to Pix2Pix. GMCNN is still the architecture for GAN, but its generators use a multi-scale CNN structure. In terms of feature extraction, the use of a multicolumn structure allows the image to be decomposed into components with different receptive fields, synthesizing features at different scales. The key problem with GMCNN is that it does not apply to large-scale datasets. Using a wavelet-based inpainting technique, WaveFill divides images into three frequency bands and fills in the areas that are absent in each band independently. This method effectively mitigates the conflict between low and high frequencies. That's why it's better in painting than the previous 3 methods, e.g., the complex structure of the mouth visually looks more coherent. In contrast, within 1000 epochs with more accurate local texture information, our model can reconstruct symmetrical, full-face images. In rows 1-3, for instance, the mouth texture is distinct and symmetrical.

To further verify that our proposed model can restore semantically informative face occlusion images, we also try to add rectangular occlusions to the face images, where the length and width of the occlusion region are 1/2 of the length and width of the images, respectively. The experimental results of the comparison are shown in Figure 9. The inpainting effect of our improved model is better than the other four methods. Compared with Figure 8, although the area of the rectangular occlusion region is larger than the face mask occlusion region, we find that the rectangular occlusion image is restored with much less blurring and artifacts. This is because the rectangular occlusion is structurally symmetric and fixed in size, whereas the face masks have different sizes, traits, and angles as they are added to the faces. The inpainting results are visually more coherent and natural. But compared to the original images in column (a), the larger areas of occlusions lead to greater diversity, the restored images are less similar to the original images.

2) Quantitative analysis

To objectively compare the inpainting results of the improved model with other methods, the inpainting outcomes are statistically evaluated in this paper using three indicators: peak signal-to-noise ratio (PSNR) [34], structural similarity (SSIM) [35] and learned perceptual image patch similarity (LPIPS) [36]. PSNR assesses the restored effect by comparing the two samples' pixel values. SSIM analyzes the variations in brightness, contrast, and structure. LPIPS is a deep learning-based image similarity assessment metric that evaluates the similarity of two images by comparing their local perceptual features. Within the realm of image inpainting, they are three of the most commonly used evaluation metrics.

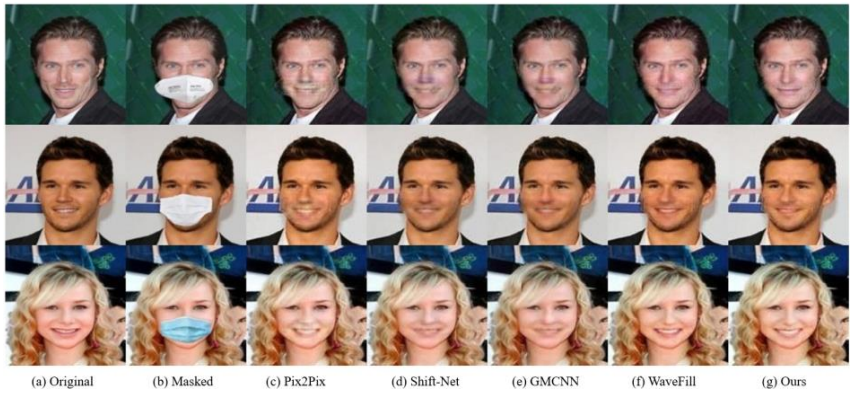


Fig. 8. Visual comparison of our model with other models on the face mask occlusion dataset. From left to right: (a) original images without masks, (b) input masked images, (c) Pix2Pix, (d) Shift-Net, (e) GMCNN, (f) WaveFill, (g) outputs of our model.

Fig. 9. Visual comparison of our model with other models on the rectangular occlusion dataset. From left to right: (a) original images without masks, (b) input masked images, (c) Pix2Pix, (d) Shift-Net, (e) GMCNN, (f) WaveFill, (g) outputs of our model.

The quantitative comparison results for the face mask occlusion dataset are shown in Table II. For image inpainting, the larger their value, the better the inpainting result is. Table III shows comparisons of quantitative results on the rectangular occlusion dataset. Experimental results display that our proposed face inpainting model is superior to several other methods. We also find that the experimental results of the rectangular occlusion face inpainting are numerically worse than the mask occlusion face. This is also in line with our previous analysis, as the rectangular occlusion covers most of the face area. The restored image is enough to ensure the coherence and naturalness of the restored face image. While PSNR, SSIM and LPIPS mainly evaluate the similarity between the restored image and the original image, Table III is numerically lower than Table II.

TABLE II
COMPARISONS OF QUANTITATIVE RESULTS ON THE FACE MASK OCCLUSION DATASET

Method	PSNR	SSIM	LPIPS
Pix2Pix	28.74	90.21	9.51
Shift-Net	29.78	91.12	9.14
GMCNN	29.57	90.61	9.15
WaveFill	30.19	91.24	8.93
Ours	30.75	91.96	8.13

TABLE III
COMPARISONS OF QUANTITATIVE RESULTS ON THE RECTANGULAR OCCLUSION DATASET

Method	PSNR	SSIM	LPIPS
Pix2Pix	24.99	85.83	13.32
Shift-Net	25.34	86.16	13.14
GMCNN	24.40	83.11	16.59
WaveFill	25.41	86.72	12.43
Ours	26.39	88.24	11.00

3) Ablation experiments

To verify the validity of our model structure, we performed ablation experiments. In the complete generator network model, spatial attention in MMRM, channel attention in MMRM, the entire MMRM module, and spatial attention in up-sampling are removed, respectively. The ablation experiments are then performed on the face mask occlusion dataset and the rectangular occlusion dataset, as shown in Figure 10 and Figure 11.

MMRM can effectively perceive global structural and semantic information to reconstruct complete edge information and improve edge distribution in occluded regions. For example, in Figure 10(e) and Figure 11(e), there are prominent boundaries and blurring. The generator model utilizes the convolutional branches in the MMRM module to extract the salient information in space and channel. The consistency of

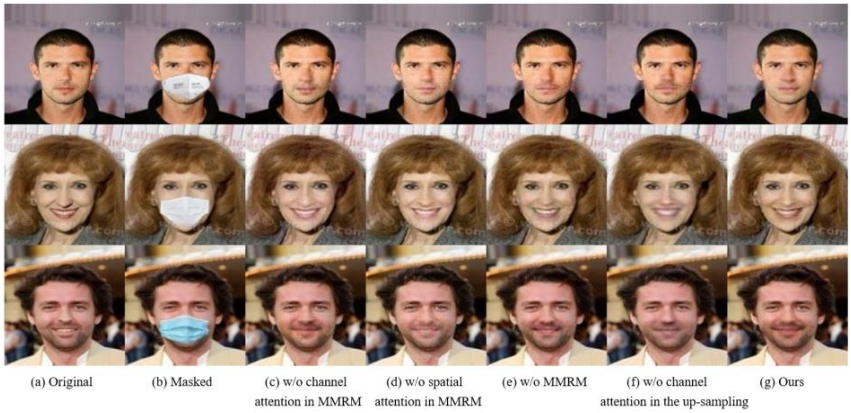


Fig. 10. Visual comparison of ablation experiments on the face mask occlusion dataset. From left to right: (a) original images without masks, (b) input masked images, (c) without channel attention in MMRM, (d) without spatial attention in MMRM, (e) without MMRM, (f) without channel attention in the up-sampling, (g) outputs of our model.

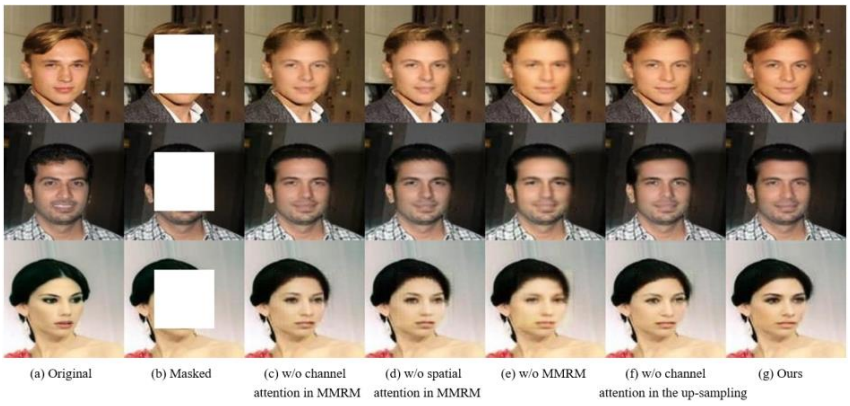


Fig. 11. Visual comparison of ablation experiments on the rectangular occlusion dataset. From left to right: (a) original images without masks, (b) input masked images, (c) without channel attention in MMRM, (d) without spatial attention in MMRM, (e) without MMRM, (f) without channel attention in the up-sampling, (g) outputs of our model.

convolutional branches in the MMRM module to extract the salient information in space and channel. The consistency of boundary information requires the gradual fusion of different spatial and channel semantics, which ultimately leads to the accurate reconstruction of global contours and local structures. Comparing columns (c) and (d), we find that if the spatial attention in MMRM is removed, there is significant blurring in the restored face image. For face inpainting, not all regions are equally important for occluded regions. With the loss functions, the spatial attention of the generator network when performing gradient backpropagation is looking for the most important parts for feature extraction. The feature maps acquired from the generator's up-sampling modules are concatenated, or superimposed at the channel level, with the left feature maps. Thus we use channel attention to focus on the key information. So, in Figure 10 (f) and Figure 11 (f), there are blurring of structurally complex parts such as the mouth and eyes.

Tables IV and V present the quantitative results of the ablation experiments. The results of the experiments on the face mask occlusion dataset and the rectangular occlusion dataset remain generally consistent with the results of the qualitative analysis above. The MMRM module in the generator model is the most useful, and once removed, the experimental data declines the most on both of the 2 datasets. The other three columns also demonstrate the role of spatial and channel attention in face image inpainting, and their removal brings about a decrease in both metrics.

TABLE IV
COMPARISONS OF QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON THE FACE MASK OCCLUSION DATASET

	w/o channel attention in MMRM	w/o spatial attention in MMRM	w/o MMRM	w/o channel attention in the up-sampling
PSNR	29.19	29.16	28.43	29.15
SSIM	91.45	91.18	90.32	91.35
LPIPS	8.37	8.61	8.75	8.44

TABLE V
COMPARISONS OF QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON THE RECTANGULAR OCCLUSION DATASET

	w/o channel attention in MMRM	w/o spatial attention in MMRM	w/o MMRM	w/o channel attention in the up-sampling
PSNR	25.86	25.12	23.45	25.74
SSIM	85.12	84.48	82.31	86.56
LPIPS	12.37	12.81	13.48	13.24

4) Cross-datasets experiments

In addition, we conducted across datasets experiments. The adopted dataset, CelebA-HQ, which is a high-quality face image designed specifically for computer vision research, is different from CelebA. Visual comparison results and quantitative comparisons are shown in Figure 12 and Table VI. The experimental results show that the inpainting effect of our design model is better than the results of several other models.

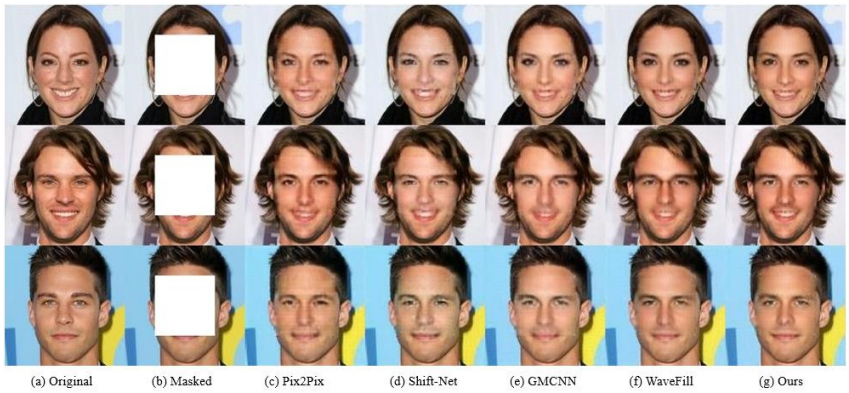


Fig. 12. Visual comparison of our model with other models on CelebA-HQ dataset. From left to right: (a) original images without masks, (b) input masked images, (c) Pix2Pix, (d) Shift-Net, (e) GMCNN, (f) WaveFill, (g) outputs of our model.

TABLE VI
COMPARISONS OF QUANTITATIVE RESULTS ON CELEBA-HQ DATASET

Method	PSNR	SSIM	LPIPS
Pix2Pix	25.28	86.26	9.78
Shift-Net	26.47	87.61	9.43
GMCNN	26.62	87.76	9.57
WaveFill	26.88	89.26	8.68
Ours	27.87	89.81	7.42

IV. CONCLUSION

We present a generative adversarial network model for face image inpainting in this paper. The discriminator network and the generator network are the two primary components of the model. The U-Net topology forms the foundation of the generator network and MMRM is used in the sampling module. This module fuses two attentional mechanisms into the residual branch, improving the network's ability to perceive global structural features and ameliorating the problem of inconsistent image inpainting boundaries. We reconstructed the dataset using the face feature point detection method and experiments demonstrate that the model performs well when reconstructing complex structures of faces. In future work, we will concentrate on reducing computational time, training the network with fewer parameters, and optimizing the architecture.

REFERENCES

[1] H. Li, W. Luo, and J. Huang, "Localization of diffusion-based inpainting in digital images," *IEEE transactions on information forensics and security*, vol. 12, no. 12, pp. 3050–3064, 2017. DOI: 10.1109/TIFS.2017.2730822.

[2] K. Li, Y. Wei, Z. Yang, and W. Wei, "Image inpainting algorithm based on TV model and evolutionary algorithm," *Soft Computing*, vol. 20, pp. 885–893, 2016. DOI: 10.1007/s00500-014-1547-7.

[3] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–10, 2012. DOI: 10.1145/2185520.2185578.

[4] T. Ružić and A. Pižurica,"Context-aware patch-based image inpainting˘ using markov random field modeling," *IEEE transactions on image processing*, vol. 24, no. 1, pp. 444–456, 2014. DOI: 10.1109/TIP.2014.2372479.

[5] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," pp. 417–424, 2000. DOI: 10.1145/344779.344972.

[6] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. IEEE, 1999, pp. 1033–1038. DOI: 10.1109/ICCV.1999.790383.

[7] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009. DOI: 10.1145/1531326.1531330.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[9] Z. Qin, Q. Zeng, Y. Zong, and F. Xu, "Image inpainting based on deep learning: A review," *Displays*, vol. 69, p. 102 028, 2021. DOI: 10.1016/j.displa.2021.102028.

[10] J. Jam, C. Kendrick, K. Walker, V. Drouard, J. G.-S. Hsu, and M. H. Yap, "A comprehensive review of past and present image inpainting methods," *Computer vision and image understanding*, vol. 203, p. 103 147, 2021. DOI: 10.1016/j.cviu.2020.103147.

[11] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Processing Letters*, vol. 51, pp. 2007–2028, 2020. DOI: 10.48550/arXiv.1909. 06399.

[12] D. J. B. Rojas, B. J. T. Fernandes, and S. M. M. Fernandes, "A review on image inpainting techniques and datasets," in *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*. IEEE, 2020. pp. 240–247. DOI: 10.1109/SIBGRAPI51738.2020.00040.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. DOI: 10.48550/arXiv.1406.2661.

[14] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544. DOI: 10.48550/arXiv.1604.07379.

[15] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017. DOI: 10.1145/3072959.3073659.

[16] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514. DOI: 10.48550/arXiv.1801.07892.

[17] C. Han and J. Wang, "Face image inpainting with evolutionary generators," *IEEE Signal Processing Letters*, vol. 28, pp. 190–193, 2021. DOI: 10.1109/LSP.2020.3048608.

[18] X. Zhang, X. Wang, C. Shi, Z. Yan, X. Li, B. Kong, S. Lyu, B. Zhu, J. Lv, Y. Yin et al., "De-gan: Domain embedded gan for high quality face image inpainting," *Pattern Recognition*, vol. 124, p. 108 415, 2022. DOI: 10.48550/arXiv.2002.02909.

[19] Y. Dogan and H. Y. Keles, "Iterative facial image inpainting based on an encoder-generator architecture," *Neural Computing and Applications*, vol. 34, no. 12, pp. 10 001–10 021, 2022. DOI: 10.48550/arXiv.2101.07036.

[20] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3911–3919. DOI: 10.48550/arXiv.1704.05838.

[21] K. Nazeri, "Edgeconnect: Generative image inpainting with adversarial edge learning," arXiv preprint arXiv:1901.00212, 2019. DOI: 10.48550/arXiv.1901.00212.

[22] J. Qin, H. Bai, and Y. Zhao, "Face inpainting network for large missing regions based on weighted facial similarity," *Neurocomputing*, vol. 386, pp. 54–62, 2020. DOI: 10.1016/j.neucom.2019.12.079.

[23] X. Yang, P. Xu, Y. Xue, and H. Jin, "Contextual feature constrained semantic face completion with paired discriminator," *IEEE Access*, vol. 9, pp. 42 100–42 110, 2021. DOI: 10.1109/ACCESS.2021.3065661.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference*, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer, 2015, pp. 234–241. DOI: 10.48550/arXiv.1505.04597.

[25] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 449–462, 2020. DOI: 10.1109/TGRS.2020.2994057.

[26] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542. DOI: 10.48550/arXiv.1910.03151.

[27] G. Chen, G. Zhang, Z. Yang, and W. Liu, "Multi-scale patch-GAN with edge detection for image inpainting," *Applied Intelligence*, vol. 53, no. 4, pp. 3917–3932, 2023. DOI: 10.1007/s10489-022-03577-2.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134. DOI: 10.48550/arXiv.1611.07004.

[30] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 1–17. DOI: 10.48550/arXiv.1801.09392.

[31] Yan Z, Li X, Li M, et al, "Shift-net: Image inpainting via deep feature rearrangement", in *Proceedings of the European conference on computer vision (ECCV)*. 2018: 1-17. DOI: 10.1007/978-3-030-01264-9_1.

[32] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," *Advances in neural information processing systems*, vol. 31, 2018. DOI: 10.48550/arXiv.1810.08771.

[33] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, and C. Miao, "Wavefill: A wavelet-based generation network for image inpainting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 114–14 123. DOI: 10.48550/arXiv.2107.11027.

[34] A. Dziembowski, D. Mieloch, J. Stankowski, and A. Grzelka, "IV-PSNR- the objective quality metric for immersive video applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7575–7591, 2022. DOI: 10.1109/TCSVT.2022.3179575.

[35] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (SSIM) revisited: A data-driven approach," *Expert Systems with Applications*, vol. 189, p. 116 087, 2022. DOI: 10.1016/j.eswa.2021.116087.

Qingyu Liu was born in 1984 in Huainan, China. He received his MSc degree (2010) in computer software and theory from Anhui Normal University, China and the PhD degree (2023) in computer science from National University, Philippine. Between 2010 and 2016, he has worked as a computer software engineer in Huainan Coal Mining (Group) Co. Ltd. He is currently a Lecturer with the School of Computer Science, Huainan Normal University, Huainan, China. His research interests include Generative Adversarial Network, image inpainting and image processing.

Lei Chen received the MSc degree (2008) in computer science from Anhui University of Science & Technology, China. He is currently a Professor with the School of Computer Science, Huainan Normal University, Huainan, China. His research interests include knowledge graph, and deep learning.

Yeguo Sun received the MSc degree (2007) and PhD degree (2010) in control science and engineering from Beijing University of Aeronautics and Astronautics, China. He is currently a Professor with the School of Finance and Mathematics, Huainan Normal University, Huainan, China. His research interests include network control systems, neural networks and finite-time control. He has published more than fifty articles indexed by SCI and EI.

Lei Liu received his MSc degree (2013) in computer science from Anhui University of Science and Technology, China and the PhD degree (2023) in computer science from National University, Philippine. He is currently an Associate Professor with the School of Computer Science, Huainan Normal University, Huainan, China. His research interests include computer vision and machine learning.

Monitoring the Semantic Change of COVID-19-related Expressions Using Dynamic Word Embeddings

Bogdán Asztalos, and Péter Bányász

***Abstract*—In this paper, we investigate how the COVID-19 pandemic has affected the use of language in the online space through measuring the semantic changes of words during the time that includes the outbreak of the pandemic and the months of the lockdown. As a first step, we apply a recent word embedding technique on a time-labelled text corpus collected from social media which represents the semantic relation of words based on their likelihood of co-occurring next to each other. By analyzing different statistical features of the received dynamic embedding, we can identify and quantitatively describe periods where the semantic properties of a chosen word are undergoing significant changes. Since this depends on the context and the usage of these words by the users, we can infer their reaction to the COVID-19-related events and relevant news dated to these periods.**

***Index Terms*—semantic change, COVID-19, social media, word embedding.**

I. INTRODUCTION

COVID-19 had a significant impact on human civilization during the last few years, inducing abrupt changes in society, ranging from the level of personal relationships and the shift towards home office in work management to the significant new challenges with which national healthcare systems had to face worldwide. Presumably, these drastic effects have affected the way we think about many things including interpersonal communication, both on personal and social levels. Reducing the number of personal contacts, online meetings becoming widespread, and several fake news popping up each week have made us change the way we approach new information and express our thoughts. Therefore, the tools of the education system, the government communication, the participants of mass media, and all the content-sharing agents also needed to adapt to these new circumstances, because – independently of their purposes – their messages will reach their goal only if they use new ways [1].

One of the most relevant examples in the topic of COVID-19 is the effectiveness of the epidemiological response of governments, which was primarily influenced by the condition of society, the degree of trust people had in their democratic institutions, the scientific community, and many other factors. The lower confidence people have in these factors, the higher the percentage of people who will refuse to accept epidemiological restrictions or vaccination in the future. However, the passage of time has led to a weakened acceptance of the government's actions. This resulted from the exhaustion of the lockdowns and the global disinformation campaigns that have been spreading on social media for years. The unclear origins of the COVID-19 virus[1] and the often-contradictory messages from different governments have provided an excellent opportunity to spread conspiracy theories and fake news. The scientific community also has a significant role in the defence against the disease, but the infodemic state related to COVID-19 has significantly eroded individuals' trust in the state of science and has reinforced pseudoscientific theories [2].

This tendency has been increased by the nature of the algorithms on social networking sites, which has led to a combination of the most absurd conspiracy theorists, from flat-earthers to anti-vaccinationists who believe that the 5G network causes the coronavirus and that microchips are being implanted in people by vaccination to track them by various actors of the deep state behind the scenes [3]. This problem is even more serious considering that the spread of fake news regarding the coronavirus and the rejection of vaccination will significantly compromise the prevention of epidemics in the future and can result more and more epidemic outbreaks globally [4]. Research has shown that social media users are more likely to believe various COVID-19-related conspiracy theories that reinforce anti-vaccination, and anti-masking, among others [5]. Well illustrates the process that in January 2020, for the first time, a post appeared on social media claiming that the spread of 5G technology was responsible for the spread of the coronavirus, which also led to a rise in anti-China feelings [6]. Then in April 2020, the 5G hoax led to 77 incidents of vandalism of 5G transmitting stations in the UK by individuals afraid of the coronavirus to prevent the further spread of the virus [7].

Therefore, it is important to counterbalance the harmful information flow in the public consciousness, and as COVID-

The research was conducted within the framework of the Network Science Research Group at Ludovika – University of Public Service. The research was supported by the EKÖP-24-3-II-ELTE-213 University Excellence Scholarship Program and the EKÖP-24-4-II-23 University Research Scholarship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

Bogdán Asztalos is with Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary, and with Health Services Management Training Centre, Semmelweis University, Budapest, Hungary (e-mail: abogdan@caesar.elte.hu)

Péter Bányász is with Department of Cybersecurity, University of Public Service Ludovika, Budapest, Hungary (e-mail: banyasz.peter@uni-nke.hu)

DOI: 10.36244/ICJ.2025.2.3

[1]In our study, we cannot and do not take a position on the origin of the coronavirus. We believe that whether it evolved naturally or was released from a laboratory is unimportant in terms of its consequences. The crucial question about the pandemic is how we can 'defeat' it.

19 itself can only be controlled through an inter- and multi-disciplinary approach, this communication-focused problem is also should examined from several directions. Such interdisciplinary cooperation is not unprecedented, similar has happened during the last 25 years, when network-based tools gradually gained relevance in exploring the mechanisms and principles of complex systems [8], including both biological [9] and social systems [10], and network thinking became accepted both by academia [11], industry [12], and governments [13]. The spreading use of artificial intelligence (AI) tools that we see today in all areas of life also highlights that computational techniques are quite helpful and over time they will be indispensable in fully understanding both social, scientific, and governmental problems [14]–[16].

The increasing prevalence of fake news and counterfeit profiles on social media platforms poses a substantial threat to informational integrity and societal trust. Whereas conventional approaches, such as rule-based solutions, metadata analysis, or human fact-checking, have assisted in mitigating the spread of disinformation, they often prove inadequate against the more sophisticated and rapidly disseminating false content [17]. Recent studies indicate that natural language processing (NLP) techniques are more effective in the identification of disinformation than traditional methods relying on fact-checking and source evaluation as they facilitate the representation and analysis of textual data. This enhances the accuracy of detection and fosters a more nuanced understanding of the semantic meanings embedded within the conveyed information [18], [19]. Such methodologies enable meticulous detection and elimination of propaganda while preserving document-level coherence, grammatical integrity, and the content's authenticity in news articles [20]. It is also found that the integration of NLP with machine learning (ML) algorithms can yield effective results in identifying false news and profiles [17], [21], [22].

Among the tools of NLP, word embedding methodologies have turned out one of the most instrumental in the identification of disinformation disseminated on social media platforms as they facilitate a deeper comprehension of the context and semantics of text [23], [24]. Word embedding algorithms like GloVe, TF-IDF, Word2Vec, and FastText-based embeddings provide sophisticated text analysis and classification techniques [25], and have proven effective in detecting disinformation within social media contexts [21], [26]. In the domain of deep learning integration, neural network models leverage embeddings to effectively capture contextual and sequential information inherent in textual data, thus markedly enhancing the accuracy of fake news classification, and understanding of context and the structural elements of disinformation [27], [28].

Besides, word embeddings provide a quantitative method to define semantic space and measure semantic distance i.e. the degree of difference between meanings of different words. By using temporal data, word embedding is also capable of measuring the semantic changes of words and gives us an opportunity to observe what phenomena have happened in human language [29], [30]. As we show in this paper, this could be used for finding out how COVID-19 has affected the way people use language, and so indirectly identifying social events and processes that can be intervened with the right communication and decision-making.

Our study aims to engage in the global scientific discourse about the problems that the pandemic created with our humble tools, proposing a new methodology that we believe can be adopted by different disciplines.

In this paper, firstly we briefly overview the necessary notions of distributional semantics and the Word2vec word embedding method, and then we introduce our results about applying diachronic word embedding on textual data from online social data.

II. Distributional semantics and word embedding

Distributional semantics (DS) is a subfield of computational linguistics that quantifies semantic similarities and differences between words and linguistic terms [31]. Since the scope of our study is to conduct a quantitative investigation about how much the meaning of words has changed during the COVID-19 pandemic, we relied heavily on the concepts of DS. The main tool we used is word embedding which represents words mathematically in a high dimension space and hence allows us to consider them as objects in a real geometric space. However, to understand word embedding and the logic behind constructing it, one should be aware of the distributional hypothesis and some of its consequences.

A. Distributional hypothesis and the role of co-occurrence statistics

It is a well-known fact that knowing the dictionary definition of a word and using it in a real-life sentence are not equivalent. The former is not even required for the latter, however, the way people use words highly depends on their lexical meaning, so it is possible to infer the semantic information from the linguistic context [32], [33]. This principle led Harris to the following statement: "a word is characterized by the company it keeps", which later became known as the distributional hypothesis and now is the basis of DS [34], [35].

One of the consequences of the distributional hypothesis is the fact that the statistical properties of linguistic context encode the meaning of words. Hence, a computational analysis of word occurrences in a large text corpus will let us distinguish different word meanings from each other and recognize similarities and differences between them [36]. For example, the word *dog* probably has more common neighbouring words with *cat* than with *house*, because there are more sentences where *dog* and *cat* are interchangeable than *dog* and *house* are. This simply follows from the fact that the meaning of the word *dog* is closer to the meaning of *cat* than to the meaning of *house*.

Hence, to express the semantic content of a word computationally, one needs to collect the words occurring in its context and analyse their statistical distribution. To store such statistics, the simplest solution is to count the co-occurrences of word pairs, so to gain the statistics of one individual word, only the concerning co-occurrence data values need to be accounted for. Therefore, studies in computational linguistics

focus on the co-occurrence statistics of text corpora and languages [37], [38].

Collecting co-occurrence statistics allows us to perform many computational analyses. One of the most apparent techniques is to build co-occurrence networks by linking words occurring in each other's context. Such a network should store all the valuable relational information about linguistic patterns, and by examining it explicitly with the tools of network study, it can reveal the fundamental semantic relationships [39], [40]. Observing the evolution of these networks can unfold the language evolution which is responsible for the network structure behind the scenes. In our study, we have focused on word embedding, which is obtained by a more complex derivation from co-occurrence statistics, but the above close relationship between co-occurrence networks and embeddings can be used both in the fields of computational linguistics and network science [41].

B. Word embedding and Word2vec

Word embedding is a technique that represents the above abstract concept of "co-occurrence statistics" and makes the semantic meaning interpretable for computers. The idea is to assign a set of real numbers (i.e., a multidimensional vector) to words in such a way that the algebraic properties of vectors reflect the same relationships between them as those that exist between the words. For instance, if two words have similar meanings, their vector representation should be close to each other in the multidimensional space, called the embedding space. Therefore, word embedding simply transforms semantic differences into geometric distances [42].

Constructing this vector representation can be done in many ways. The simplest methods, like LSA, take the co-occurrence value of some pre-selected context words and apply some kinds of dimension reduction on them [43], but in the last decade, more sophisticated word embedding techniques were introduced using unsupervised ML algorithms [44].

During our investigation, we used Word2vec's Skip-gram model with negative sampling (SGNS), which is one of the most widely used word embedding methods in recent years. It also uses unsupervised learning: the objective of the model is to estimate a co-occurrence probability using a two-layer neural network and the vector representation of the words is a kind of by-product of the optimization process. The cause of its popularity is that it can process a large amount of data in a short time because of its sampling procedure and also the fact that the constructed embedding returns psycho-linguistical relations more efficiently than other alternatives [45], [46]. The output of such an embedding is visualised in Fig. 1.

III. METHODS

A. Gathering linguistic data with SentiOne

To receive data for the linguistic study, we used an AI-powered social listening tool, called SentiOne with which we searched for COVID-19-related words and expressions, and downloaded posts and web pages containing them. SentiOne is a content-based web analytics platform that is dedicated to crawling and analysing content perceived online channels, like social media platforms, news portals, blogs, etc. [47]. It covers and recognizes 70 languages across the globe. The platform monitoring tool currently monitors over 20 000 000 000 mentions and gathers data from 8 different types of sources, namely portals, blogs, Twitter, Facebook, Instagram, video, forums, and review sites.

The mentions are divided into statements and articles, statements being automatically classified as either positive, neutral, or negative with the use of SentiOne's unique, proprietary algorithm. The platform's sentiment analysis is based on research work by John R. Crawford and Julie D. Henry [48]. They analysed the Positive and Negative Affect Schedule (PANAS). Based on their research, SentiOne's developers created algorithms that help determine the author's emotional attitude to the discussed topic. The platform uses proprietary artificial intelligence algorithms to classify the posts' overall sentiment.

The interactive platform is built upon user-provided keywords and key phrases to look for the specific mentions that, either in themselves or within their context, contain those pre-given phrases that interest the user. The system gathers data in almost real-time yet has a memory that can go back up to 3 years. For quantitative research, data is structured by different focus points and research parameters and is visualized interactively. This technology also supports qualitative research, enabling in-depth analysis and categorizing all the indexed web content.

We used three groups of keywords and key expressions to capture social data: some common words (mostly names of professions) that are not related to COVID-19 in an obvious way and can serve as a stable meaning benchmark; the names of the different vaccine types; and an expression aimed at the 5G-related fake news. The used keywords and expressions are listed in Table I. We have searched for each keyword in 30 different languages[2]. In this research, these queries were run because they are language-independent and have the same meaning in each of the given languages.

B. Diachronic word embedding

To measure how much words change their meaning over time, we used a series of word embedding. The steps of the process with which we received these embeddings is illustrated in Fig. 2.

Since the input of an embedding method is the co-occurrence data as explained in Section II-A, it is possible to construct distinct embeddings for each month starting from the collected statistics corresponding to the individual months. The result of a single embedding is a set of word vectors (i.e. points in a high-dimensional space) representing the meaning relations of the words, so if we get different sets for different months, we can compare them, hence inferring what changed in the language during a month. We assume that

[2]Currently, the following languages are available in fully supported form: Polish, German (+ Swiss German, Austrian German), Russian, Ukrainian, English (+ UK, US, Ireland), Dutch (+ Belgian Dutch), French (+ Belgian French, Swiss-French), Slovenian, Slovak, Hungarian, Romanian, Bulgarian, Serbian, Croatian, Bosnian, Montenegrin, Czech, Danish, Finnish, Swedish, Norwegian, Latvian, Lithuanian, Italian, Spanish, Portuguese, and Greek.

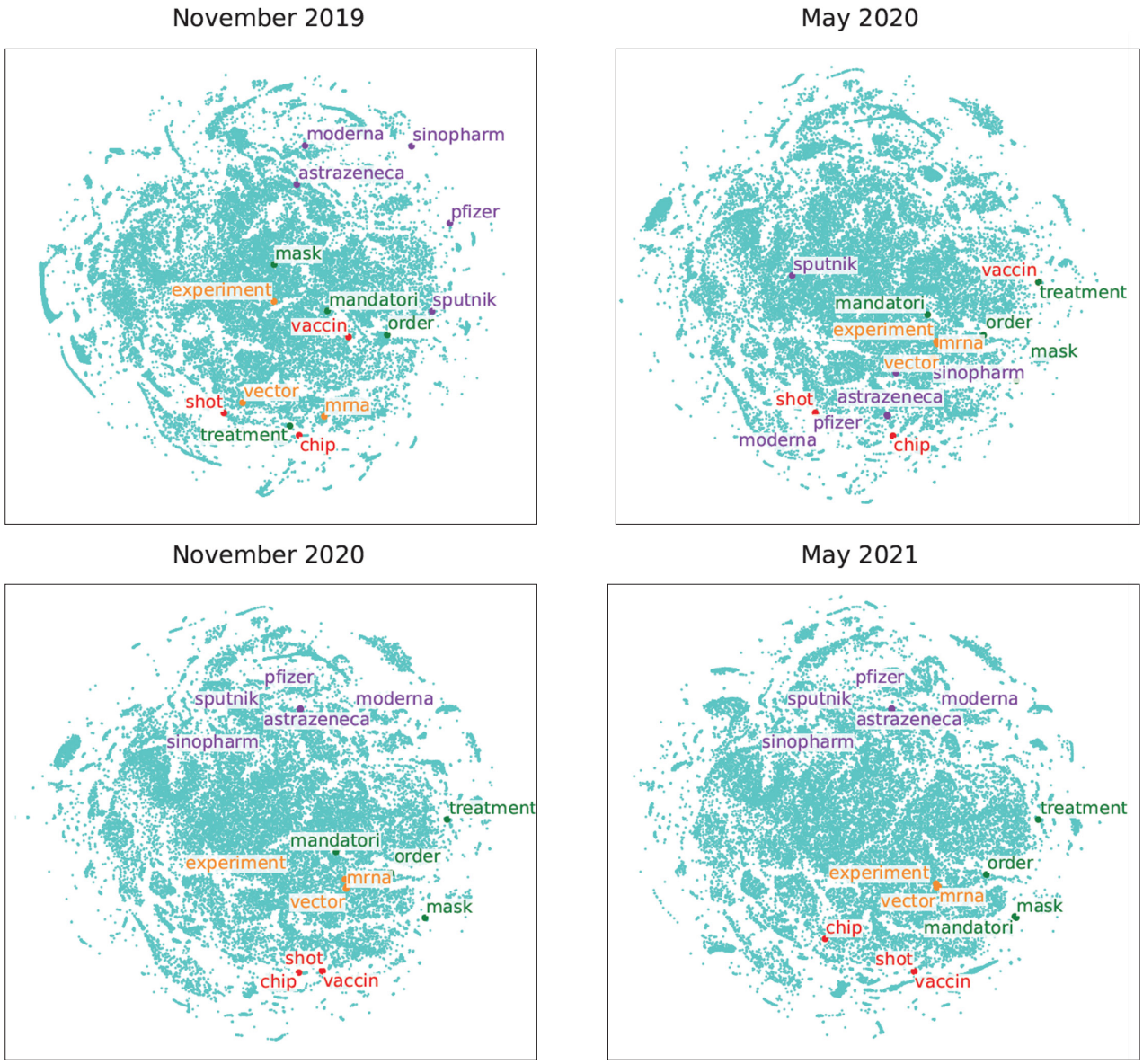


Fig. 1. Visualisation of diachronic word embedding. The coordinates of words can be considered as their position in a snapshot of the semantic space, where words are continuously walking. We plotted four different state during the examined period, where the meaning changes of the highlighted words can be observed. The geometric relation of these pandemic-related terms reflects relatively well the social attitude changes towards these topics.

most words do not change their meaning in a short time, like a month, so comparing consecutive embeddings can reflect the changes that happened so that the concerned words changed their position relative to the others. The obtained results also supported this assumption. This allows us to consider the words in the embedding space as walking particles, and to look at the semantic change in a language as the collective motion of these particles. Also, as previous research has shown, this collective motion of words can be studied quantitatively, and their behaviour can be described as an anomalous diffusion with a significant subdiffusive character which might be derived from the governing non-linear microscopic dynamics of human language evolution [49], [50].

To build such a series of embedding, we have run through all the textual sources in the downloaded raw data and performed the data processing methodology published by Asztalos et al. [49]. The used embedding method was Word2vec's Skip-gram model with negative sampling (SGNS) [45], [46] into 300 dimensions.

Since the dataset contained time-labels for each post about when they were published, we could distinguish different co-occurrence statistics coming from different periods. We applied monthly separation, i.e., handled independently the statistics from the 24 months between July 2019 and June 2021. The received series of embedding can be interpreted as snapshots of how word meanings migrate over the overall cloud of words. Four of these snapshots are illustrated in 2D in Fig. 1.

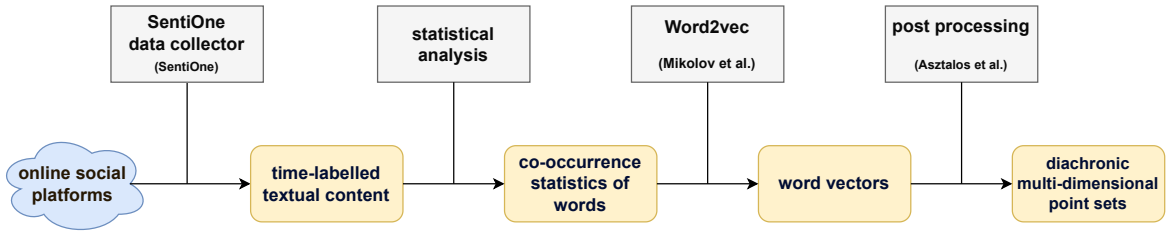


Fig. 2. The steps of our data process flow.

TABLE I
LIST OF KEYWORDS AND EXPRESSIONS WE USED FOR SEARCHING

Common words	Vaccine names	Fake news
"Bread"	"AstraZeneca"	"5G" and "COVID"
"Driver"	"J&J"	
"Engineer"	"Johnson and Johnson"	
"Film director"	"Moderna"	
"Illness"	"Pfizer"	
"Journalist"	"Sinopharm"	
"Lawyer"	"Sinovac"	
"Miner"	"Sputnik V"	
"Principal"		
"Soldier"		

IV. RESULTS

While we have analyzed the spread of fake news based on several keywords, we decided to focus on the trend of spreading COVID-19 and 5G-related fake news, which led to mass destruction in the UK, as it was mentioned in the Introduction. For the search terms "COVID" and "5G", we identified 1,156,535 shares during the period under study, resulting in approximately 4.4 billion views

As can be seen in panel (a) in Fig. 3, the significant majority of shares are found to be neutral, which can be attributed to three factors:

- AAI examines the shared content based on context, evaluating the expressions it contains in terms of positive and negative aspects, and then deciding which sentiment to attach to it based on the results. However, if the number of positive and negative expressions is approximately equal, the algorithm labels it as neutral.
- In many cases, users simply reshare the content without any explanation or comment, so they cannot analyse the new context.
- A more recent emerging new phenomenon is the intentional use of terms and phrases in the content shared for the express purpose of disinformation, overloading or poisoning the meaning of words, and deliberately creating a different emotional context. For example, the hashtag "#Bidenworstpresidentever" might be used as a positive hashtag in trending posts, even though the hashtag's meaning clearly indicates a negative emotional connotation.

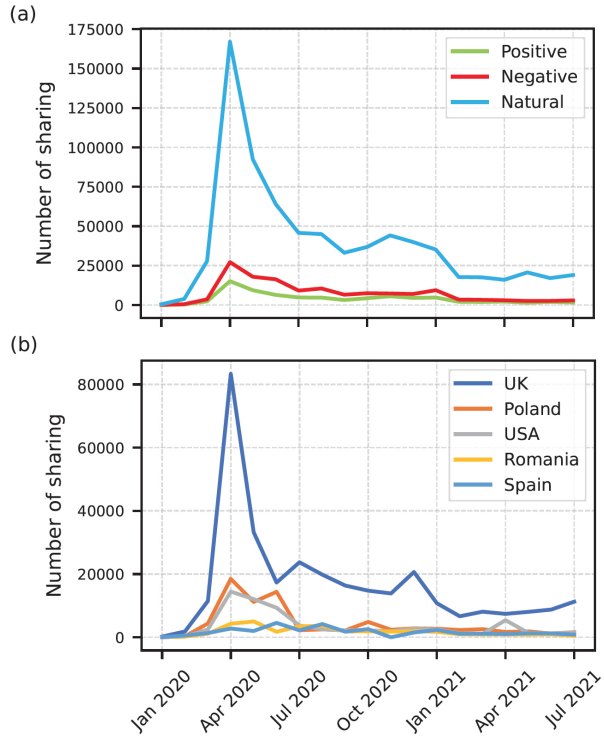


Fig. 3. Number of shares over time found by the keyword "COVID" and "5G". In panel (a) they are grouped by sentiment, while in panel (b) they are grouped by the country.

Fig. 3 shows that the number of shares related to the terms "COVID" and "5G" significantly increased from March 2020, reaching a record 230,000 shares in April, which coincides with the previously cited destruction of 77 5G towers. If we compare this with the territorial spread of the sharing (see Fig. 3, panel (b)), it is not surprising that the dominant location of the sharing is the UK: overall, there are 317,313 shares for the two search terms above, which, narrowed down to April 2020, means more than 80,000 shares.

Concrete examples of how we can visualize linguistic changes with the help of word embeddings can be seen in Fig. 1. Word embedding methods (like Word2vec) place words in a high-dimensional space so that words with similar meanings are closer to each other. Typically, the number of dimensions is several hundred (in our case 300), so the procedure result is relatively abstract. Still, with dimension reduction tools, it is possible to illustrate the structure of the "word cloud" in 2D. Using the t-SNE visualization [51], we plotted four

reaching a record 230,000 shares in April, which coincides with the previously cited destruction of 77 5G towers. If we compare this with the territorial spread of the sharing (see Fig. 3, panel (b)), it is not surprising that the dominant location of the sharing is the UK: overall, there are 317,313 shares for the two search terms above, which, narrowed down to April 2020, means more than 80,000 shares.

Concrete examples of how we can visualize linguistic changes with the help of word embeddings can be seen in Fig. 1. Word embedding methods (like Word2vec) place words in a high-dimensional space so that words with similar meanings are closer to each other. Typically, the number of dimensions is several hundred (in our case 300), so the procedure result is relatively abstract. Still, with dimension reduction tools, it is possible to illustrate the structure of the "word cloud" in 2D. Using the t-SNE visualization [51], we plotted four states of the word cloud in Fig. 1. I.e., these figures are approximations of the high-dimensional objects representing semantical meanings. It can be seen that the overall structure of the cloud (i.e., general meaning relations in the language) is the same (supporting our basic assumption), but some pandemic-related words get closer to each other while some get further. Thus, it can be observed, for instance, how the different pharmaceutical companies manufacturing vaccines grouped together, or how the word *chip* was firstly associated with vaccines and moved away later.

To perform some more descriptive analysis, we collected the 30 words closest to some chosen test words each month and plotted the monthly changes in this set in Fig. 4. Suppose that the spatial distance represents the meaning difference between words. In that case, we can assume that a word changes its closest neighbours because its meaning changes over time, so the number of changed neighbours correlates with its meaning change, hence can be viewed as a "rate of meaning change". In Fig. 4 this value is plotted over time for four everyday words (panel (a)) and for four vaccine names (panel (b)). One can see those words with stable, conventional meanings (e.g., word, friend) have a lower change rate of around 10-15; this can be considered the static value. In contrast, words like mask or president are more agile in the scanned period, probably because of the social events that happened that time. Vaccine words (more precisely the names of the companies producing them) had no fixed neighbour set in the beginning, but they decayed to the static value when the collective social opinion about them reached its final state so that we can identify the acceptance order of different vaccines, consistent with real-life experience.

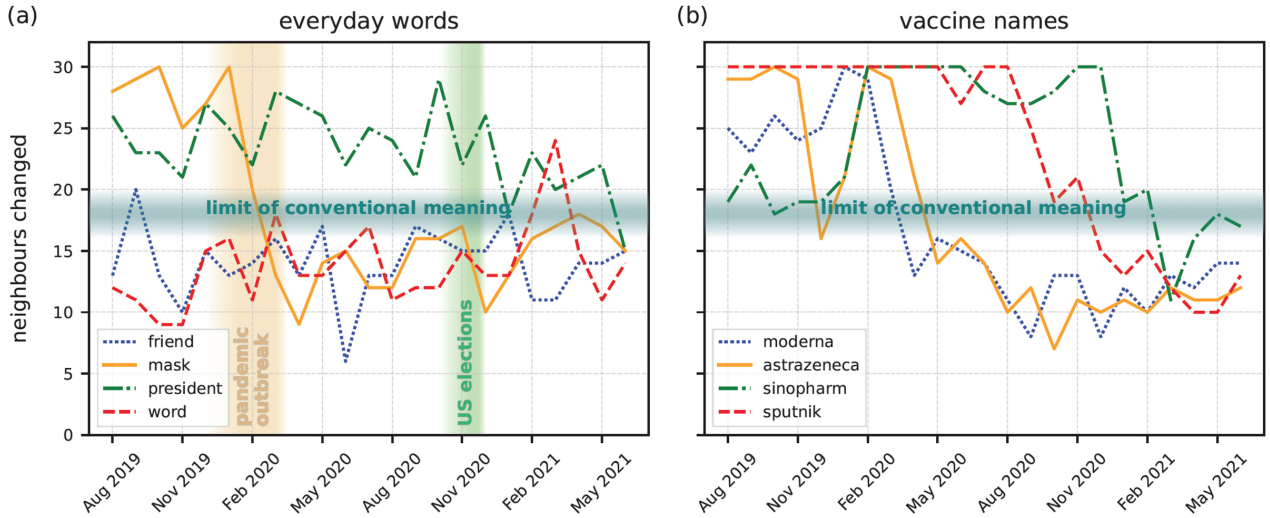


Fig. 4. The number of neighbours (out of 30) of a certain word which change from one month to another. Panel (a) shows four everyday words while panel (b) shows four vaccine words. Meaning of words with lower value than 16-20 can be viewed as static. On panel (a), the effect of social events like the outbreak of COVID or the US election campaign can be identified, while on panel (b), the gradual acceptance of vaccines can be observed.

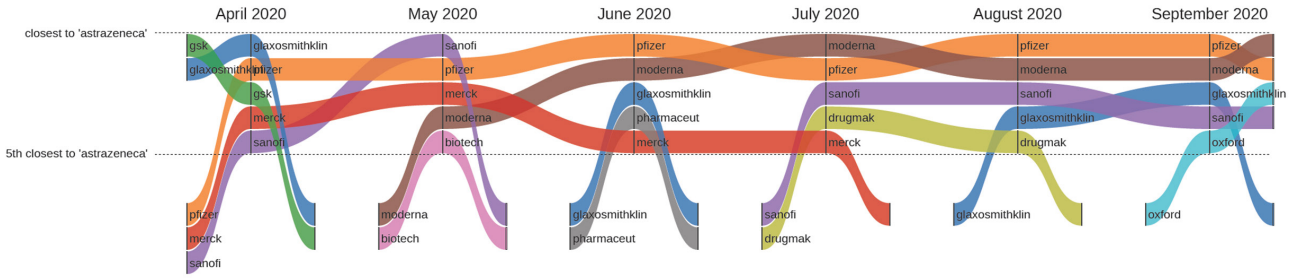


Fig. 5. The five closest words to AstraZeneca by time. The upmost stripe represents the closest word, while the lowest represents the 5th closest. The social change in what terms is associated with AstraZeneca over time is well observable.

To study in more detail what happens in the studied period, we can look for the closest few words to a certain word each month. Fig. 5 presents the case of AstraZeneca for its 5 closest neighbours. This Sankey figure shows the change of these words and their order over time. The following change in social attitude over time can observed: Pfizer and Moderna became really close when they started to develop the COVID-19 vaccine, but Merck and Glaxo Smith Kline (which are also pharmaceutical companies but did not have their own vaccine) left the immediate neighborhood. This also underlines the fact that public opinion about AstraZeneca (especially on social media platforms) was influenced primarily by the vaccine-manufacturing activity of the company in the examined period.

V. CONCLUSION

The results of this study hold significant implications for both scientific inquiry and practical strategies aimed at combating disinformation, particularly in relation to the evolving nature of fake news and linguistic shifts observed during the COVID-19 pandemic. The implementation of diachronic word embedding technologies has established a novel methodological framework for tracing the temporal development of linguistic meanings, thereby enabling an analysis of the transformation of social narratives and emotional patterns.

A key innovative aspect of the methodology is the application of linguistic representations within an interdisciplinary framework, facilitating not only textual analyses but also a comprehensive examination of the dynamic phenomena associated with social change. The findings can inform the formulation of targeted strategies to combat disinformation and enhance the efficacy of political and public health communications. Investigating the evolving context surrounding words linked to fake news will illuminate particularly virulent topics and phrases circulating on social media, thus allowing for more effective deployment of early intervention measures.

Furthermore, the visualization of linguistic embeddings facilitates both qualitative and quantitative assessments of shifts in social narratives. This dual methodological framework strengthens interdisciplinary research by integrating insights from network theory, NLP, and emotional analysis. The incorporation of new techniques and methodologies is poised to address pressing global challenges, such as epidemics and political conflicts, at both theoretical and practical levels.

ACKNOWLEDGMENT

The authors thank Péter Pollner and Gergely Palla for useful discussions at various stages of the project, and for the help with many technical details.

REFERENCES

[1] T. Bat-Erdene, N. Zayed Yazan, X. Qiu, I. Shakoor, A. Mekni, P. A. Kara, L. Bokor, and A. Simon, "On the quality of experience of content sharing in online education and online meetings," *Infocommunications Journal*, vol. 14, no. 2, pp. 73–84, 2022. **DOI**: 10.36244/ICJ.2022.2.8

[2] J. Hua and R. Shaw, "Corona virus (covid-19) "infodemic" and emerging issues through a data lens: The case of china," *International journal of environmental research and public health*, vol. 17, no. 7, p. 2309, 2020. **DOI**: 10.3390/ijerph17072309

[3] D. Allington, B. Duffy, S. Wessely, N. Dhavan, and J. Rubin, "Healthprotective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency," *Psychological medicine*, vol. 51, no. 10, pp. 1763–1769, 2021. **DOI**: 10.1017/S003329172000224X

[4] I. Ullah, K. S. Khan, M. J. Tahir, A. Ahmed, and H. Harapan, "Myths and conspiracy theories on vaccines and covid-19: Potential effect on global vaccine refusals," *Vacunas*, vol. 22, no. 2, pp. 93–97, 2021. **DOI**: 10.1016/j.vacun.2021.01.001

[5] D. Romer and K. H. Jamieson, "Patterns of media use, strength of belief in covid-19 conspiracy theories, and the prevention of covid-19 from march to july 2020 in the united states: survey study," *Journal of medical Internet research*, vol. 23, no. 4, p. e25215, 2021. **DOI**: 10.2196/25215

[6] M. Rich, "As coronavirus spreads, so does anti-chinese sentiment," *The New York Times*, Jan 2020, available at: https://www.nytimes.com/2020/01/30/world/asia/coronavirus-chinese-racism.html, accessed: 12 Jun 2023.

[7] C. Reichert, "5g coronavirus conspiracy theory leads to 77 mobile towers burned in uk, report says," https://www.cnet.com/health/5g-coronavirus-conspiracy-theory-sees-77-mobile-towers-burned-report-says/, 2020, accessed: 12 Jun 2023.

[8] M. Mitchell, "Complex systems: Network thinking," *Artificial intelligence*, vol. 170, no. 18, pp. 1194–1212, 2006. **DOI**: 10.1016/j.artint.2006.10.002

[9] E. Alm and A. P. Arkin, "Biological networks," *Current opinion in structural biology*, vol. 13, no. 2, pp. 193–202, 2003. **DOI**: 10.1016/S0959-440X(03)00031-9

[10] J. Scott, *Social networks: Critical concepts in sociology*. Taylor & Francis, 2002, vol. 4.

[11] D. J.Watts, "The "new" science of networks," *Annu. Rev. Sociol.*, vol. 30, no. 1, pp. 243–270, 2004. **DOI**: 10.1146/annurev.soc.30.020404.104342

[12] B. Axelsson and G. Easton, *Industrial networks (routledge revivals): A new view of reality*. Routledge, 1992.

[13] S. J. Ball and C. Junemann, Networks, new governance and education. *Policy Press Bristol*, 2012.

[14] S. Makridakis, "The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms," *Futures*, vol. 90, pp. 46–60, 2017. **DOI**: 10.1016/j.futures.2017.03.006

[15] M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam et al., "On scientific understanding with artificial intelligence," *Nature Reviews Physics*, vol. 4, no. 12, pp. 761–769, 2022. **DOI**: 10.1038/s42254-022-00518-3

[16] V. J. Straub, D. Morgan, J. Bright, and H. Margetts, "Artificial intelligence in government: Concepts, standards, and a unified framework," *Government Information Quarterly*, vol. 40, no. 4, p. 101 881, 2023. **DOI**: 10.1016/j.giq.2023.101881

[17] E. Papageorgiou, C. Chronis, I. Varlamis, and Y. Himeur, "A survey on the use of large language models (llms) in fake news," *Future Internet*, vol. 16, no. 8, p. 298, 2024. **DOI**: 10.3390/fi16080298

[18] R. Kozik, G. Ka ̨tek, M. Gackowska, S. Kula, J. Komorniczak, P. Ksieniewicz, A. Pawlicka, M. Pawlicki, and M. Chorás, "Towards explainable fake news detection and automated content credibility assessment: Polish internet and digital media use-case," *Neurocomputing*, vol. 608, p. 128 450, 2024. **DOI**: 10.1016/j.neucom.2024.128450

[19] S. E. V. S. Pillai, "Enhancing misinformation detection through semantic analysis and knowledge graphs," in *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*. IEEE, 2024, pp. 1–5. **DOI**: 10.1109/ICDECS59733.2023.10503553

[20] A. Bhagat, F. Mallick, N. Karia, and A. Kaushal, "Indeprop: Information-preserving de-propagandization of news articles (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 915–12 916. **DOI**: 10.1609/aaai.v36i11.21594

[21] H. Kaur, "Fake news detection using semantic analysis and machine learning techniques," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2023, pp. 1–6. DOI: 10.1109/ICCCNT56998.2023.10307799

[22] A. A. Ali, S. Latif, S. A. Ghauri, O.-Y. Song, A. A. Abbasi, and A. J. Malik, "Linguistic features and bi-lstm for identification of fake news," *Electronics*, vol. 12, no. 13, p. 2942, 2023. DOI: 10.3390/electronics12132942

[23] P. Mitra and L. Jacob, "Fake news detection and classify the category," in *2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*. IEEE, 2022, pp. 1–7. DOI: 10.1109/TQCEBT54229.2022.10041596

[24] C.-O. Truiča and E.-S. Apostol, "It's all in the embedding! fake news detection using document embeddings," *Mathematics*, vol. 11, no. 3, p. 508, 2023. DOI: 10.3390/math11030508

[25] A. Neelima and S. Mehrotra, "A comprehensive review on word embedding techniques," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*. IEEE, 2023, pp. 538–543. DOI: 10.1109/ICISCoIS56541.2023.10100347

[26] M. A. Al-Tarawneh, O. Al-irr, K. S. Al-Maaitah, H. Kanj, and W. H. F. Aly, "Enhancing fake news detection with word embedding: A machine learning and deep learning approach," *Computers*, vol. 13, no. 9, p. 239, 2024. DOI: 10.3390/computers13090239

[27] S. Pande, S. Rathod, R. Joshi, G. Chvan, D. Jadhav, P. Phutane, S. Gonge, and K. Kadam, "Fake news identification using regression analysis and web scraping," *IJSSE*, vol. 12, pp. 311–318, 2022. DOI: 10.18280/ijsse.120305

[28] S. A. Althubiti, F. Alenezi, and R. F. Mansour, "Natural language processing with optimal deep learning based fake news classification." *Computers, Materials & Continua*, vol. 73, no. 2, 2022. DOI: 10.32604/cmc.2022.028981

[29] P. J. Worth, "Word embeddings and semantic spaces in natural language processing," *International journal of intelligence science*, vol. 13, no. 1, pp. 1–21, 2023. DOI: 10.4236/ijis.2023.131001

[30] K. Durrheim, M. Schuld, M. Mafunda, and S. Mazibuko, "Using word embeddings to investigate cultural biases," *British Journal of Social Psychology*, vol. 62, no. 1, pp. 617–629, 2023. DOI: 10.1111/bjso.12560

[31] G. Boleda, "Distributional semantics and linguistic theory," *Annual Review of Linguistics*, vol. 6, no. 1, pp. 213–234, 2020. DOI: 10.1146/annurev-linguistics-011619-030303

[32] W. E. Nagy, P. A. Herman, and R. C. Anderson, "Learning words from context," *Reading research quarterly*, pp. 233–253, 1985. DOI: 10.2307/747758

[33] C. Van Petten, "Words and sentences: Event-related brain potential measures," *Psychophysiology*, vol. 32, no. 6, pp. 511–525, 1995. DOI: 10.1111/j.1469-8986.1995.tb01228.x

[34] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954. DOI: 10.1080/00437956.1954.11659520

[35] A. Lenci et al., "Distributional semantics in linguistic and cognitive research," *Italian journal of linguistics*, vol. 20, no. 1, pp. 1–31, 2008.

[36] C. P. Cook, *Exploiting linguistic knowledge to infer properties of neologisms*. University of Toronto, 2010.

[37] P. Edmonds, "Choosing the word most typical in context using a lexical co-occurrence network," *arXiv preprint cs/9811009*, 1998. DOI: 10.48550/arXiv.cs/9811009

[38] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behavior research methods*, vol. 39, pp. 510–526, 2007. DOI: 10.3758/BF03193020

[39] K. Stuart and A. Botella, "Corpus linguistics, network analysis and cooccurrence matrices," *International Journal of English Studies*, vol. 9, no. 3, pp. 1–20, 2009.

[40] J. Cong and H. Liu, "Linguistic emergence from a networks approach: The case of modern chinese two-character words," *Plos one*, vol. 16, no. 11, p. e0259818, 2021. DOI: 10.1371/journal.pone.0259818

[41] G. Budel, Y. Jin, P. Van Mieghem, and M. Kitsak, "Topological properties and organizing principles of semantic networks," *Scientific Reports*, vol. 13, no. 1, p. 11728, 2023. DOI: 10.1038/s41598-023-37294-8

[42] S. S. Birunda and R. K. Devi, "A review on word embedding techniques for text classification," in *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*. Springer, 2021, pp. 267–281. DOI: 10.1007/978-981-15-9651-3_23

[43] S. T. Dumais, "Latent semantic analysis," *Annual Review of Information Science and Technology (ARIST)*, vol. 38, pp. 189–230, 2004. DOI: 10.1002/aris.1440380105

[44] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 298–307. DOI: 10.18653/v1/D15-1036

[45] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[46] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013. DOI: 10.48550/arXiv.1301.3781

[47] SentiOne, "Social listening 101," https://sentione.com/resources/sociallistening, 2023, accessed: 1 May 2023.

[48] J. R. Crawford and J. D. Henry, "The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample," *British journal of clinical psychology*, vol. 43, no. 3, pp. 245–265, 2004. DOI: 10.1348/0144665031752934

[49] B. Asztalos, G. Palla, and D. Czégel, "Anomalous diffusion analysis of semantic evolution in major indo-european languages," *Plos one*, vol. 19, no. 3, p. e0298650, 2024. DOI: 10.1371/journal.pone.0298650

[50] D. Czégel, S. G. Balogh, P. Pollner, and G. Palla, "Phase space volume scaling of generalized entropies and anomalous diffusion scaling governed by corresponding non-linear fokker-planck equations," *Scientific reports*, vol. 8, no. 1, p. 1883, 2018. DOI: 10.1038/s41598-018-20202-w

[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Bogdán Asztalos graduated from the Eötvös Loránd University, Faculty of Science as a Physicist (M.S. degree), and currently is a Phd student in Physics. His main research interest is the mathematical background of theoretical statistical physics, but also works on social network analysis, and looks for interdisciplinary applications of the tools of statistical physics. His supervisor is Péter Pollner.

Péter Bányász graduated in political science from the Faculty of Law and Political Science of the Eötvös Loránd University. He then obtained his doctorate at the Military Engineering Doctoral School of the National University of Public Service. His dissertation is entitled "Opportunities and Risks of Social Media in the Defence Sector." His research interests include the human aspect of cyber security, network theories of psychological operations, and the relationship between privacy and surveillance. He is an assistant professor at the Faculty of Public Governance and International Studies of the University of Public Service and a researcher at the Institute for Cyber Security Research. He is also a senior mentor teacher at the University's Advanced College for National Security and Tivadar Puskás Technical College for Advanced Studies. He is also an active member of several scientific societies.

A Comparative Analysis of Static Word Embeddings for Hungarian

Máté Gedeon

Abstract—**This paper presents a comprehensive analysis of various static word embeddings for the Hungarian language, including traditional models such as Word2Vec, FastText, as well as static embeddings derived from BERT-based models using different extraction methods.**

We evaluate these embeddings on both intrinsic and extrinsic tasks to provide a holistic view of their performance. For intrinsic evaluation, we employ a word analogy task, which assesses the embeddings' ability to capture semantic and syntactic relationships. Our results indicate that traditional static embeddings, particularly FastText, excel in this task, achieving high accuracy and mean reciprocal rank (MRR) scores. Among the BERT-based models, the X2Static method for extracting static embeddings demonstrates superior performance compared to decontextualized and aggregate methods, approaching the effectiveness of traditional static embeddings.

For extrinsic evaluation, we utilize a bidirectional LSTM model to perform Named Entity Recognition (NER) and Part-of-Speech (POS) tagging tasks. The results reveal that embeddings derived from dynamic models, especially those extracted using the X2Static method, outperform purely static embeddings. Notably, ELMo embeddings achieve the highest accuracy in both NER and POS tagging tasks, underscoring the benefits of contextualized representations even when used in a static form.

Our findings highlight the continued relevance of static word embeddings in NLP applications and the potential of advanced extraction methods to enhance the utility of BERT-based models. This piece of research contributes to the understanding of embedding performance in the Hungarian language and provides valuable insights for future developments in the field. The training scripts, evaluation codes, restricted vocabulary, and extracted embeddings will be made publicly available to support further research and reproducibility.

Index Terms—**word embeddings, BERT, FastText, Word2Vec, NLP, intrinsic evaluation, extrinsic evaluation**

I. Introduction

Teaching machines to understand human language is a crucial step toward developing intelligent systems. While one-hot encoding can be effective for small-scale classification tasks, it becomes impractical when the aim is to represent thousands of words and their variations. This limitation led to the development of word embeddings—dense vector representations of words in a continuous space. These embeddings not only represent similar words with similar vectors but also are able to capture complex semantic relationships between them.

Bengio et al. [1] introduced feedforward neural networks with a single hidden layer for language modeling in 2003. These models were capable of learning distributed word representations but suffered from significant scalability issues, limiting their ability to handle large vocabularies. Collobert and Weston [2] addressed this bottleneck in 2008 by refining the training objective and demonstrating the effectiveness of word embeddings pre-trained on large corpora for downstream tasks. Their neural network architecture also served as an inspiration for many subsequent approaches.

Building on the idea that words can be effectively predicted from their context, several word embedding models emerged, including Word2Vec [3], GloVe [4], and FastText [5]. These models proved highly effective in various natural language processing (NLP) tasks. However, their primary limitation lies in their static nature: each word is assigned a single representation, making it impossible to capture context-dependent meanings (e.g., polysemy). This shortcoming paved the way for the development of dynamic, contextualized word embeddings, exemplified by Transformer-based models [6] such as BERT [7], GPT [8], and T5 [9], which generate word representations that adapt based on surrounding text.

Despite the success of dynamic models, static word embeddings remain a viable option for numerous applications due to their significantly lower computational requirements compared to context-dependent embeddings. In this paper, we analyze the performance of multiple pre-trained static word embeddings for Hungarian, as well as static embeddings extracted from BERT-based models, on both intrinsic and extrinsic tasks.

The main contributions of this paper are as follows:

- A comprehensive analysis of the performance of multiple static word embeddings for Hungarian.
- An evaluation of static embeddings extracted from BERT-based models.
- Public release of the code used in the study[1].
- Public release of the embeddings extracted from BERT-based models[2].

The rest of the paper is organized as follows: Section 2 provides an overview of related work, Section 3 presents the embedding models employed, Section 4 introduces the datasets used in the study, Section 5 describes the experiments conducted, and Section 6 summarizes findings and outlines future directions.

II. Related Work

Multiple methods exist for extracting static embeddings from BERT-based models. The simplest approach involves averaging the token embeddings of a word, when inputting only that word to the model (later referred to as *decontex-*

Máté Gedeon is with the Faculty of Natural Sciences, Budapest University of Technology and Economics, Budapest, Hungary (e-mail: gedeonm01@gmail.com).

DOI: 10.36244/ICJ.2025.2.4

[1]https://github.com/gedeonmate/hungarian_static_embeddings

[2]https://huggingface.co/gedeonmate/static_hungarian_bert

tualized method). Although easy, this method presents an input to the model, which is different from the training data, potentially leading to suboptimal embeddings. Therefore, more sophisticated methods have been proposed. In this piece of research, we employed two such methods: the *aggregate* method [10] and the *X2Static* method [11]. The *aggregate* method pools embeddings of different contexts the word appears in. This presents the model a more natural input than the *decontextualized* method, but as it uses the embedding of the whole context and not just the word itself, it still can lead to suboptimal results. The *X2Static* method, on the other hand, uses a CBOW-inspired static word-embedding approach, leveraging contextual information from a teacher model to generate static embeddings. This method has been shown to outperform the *aggregate* method in multiple tasks.

A Turkish study [12] conducted an in-depth analysis of static word embeddings in the language, which had a substantial impact on the chosen methodology in this study. Their findings were similar, finding the *X2Static* method to be the most effective for extracting static embeddings from BERT-based models.

Several studies have contributed to the evaluation of word embeddings in Hungarian. A 2019 paper [13] assessed the semantic accuracy of word embeddings through word analogy tasks, revealing a substantial performance drop of 50–75% compared to English. The authors attributed this decline to the high morphological variation in Hungarian and the less stable semantic representations that result from it.

Research on Hungarian contextual embeddings has also gained traction. A comparative study [14] evaluating huBERT against multilingual BERT models demonstrated that huBERT outperformed its multilingual counterparts in morphological probing, POS tagging, and NER tasks. On the same tasks, but with a focus on the impact of subword pooling, Ács et al. [15] conducted a comprehensive cross-linguistic analysis of pooling strategies over several languages.

III. EMBEDDING MODELS

A. Word2Vec

Word2Vec [3] is one of the most influential static word embedding models. It is based on the idea that words appearing in similar contexts have similar meanings. Unlike models focused solely on predicting the next word, Word2Vec considers both preceding and succeeding words within a fixed context window (e.g., five words). The model has two architectures: Continuous Bag of Words (CBOW) and Skip-gram. In CBOW, the goal is to predict a target word from its context, disregarding word order. In Skip-gram, the objective is to predict context words based on a given target word.

Several pre-trained embeddings exist for various languages, including Hungarian. In this study, we used pre-trained embeddings developed for the EFNILEX[3] project [16], trained on the combined Hungarian Webcorpus [17] and the Hungarian National Corpus [18] with 600 dimensions.

[3]http://corpus.nytud.hu/efnilex-vect/

B. FastText

In 2018, Bojanowski et al. [5] introduced FastText, which provides word embeddings for 157 languages. The model was trained on Wikipedia dumps and the Common Crawl. FastText follows a training procedure similar to Word2Vec but incorporates subword information. The Skip-gram model uses character n-grams, assigning a vector representation to each, and constructs word representations by summing the vectors of the character n-grams present in the word. The full word is included to maintain a unique vector for each word. The CBOW model represents words as bags of character n-grams with position-dependent weights to capture positional information. For this study, we used the 300-dimensional Hungarian embeddings published by FastText.

C. SpaCy

SpaCy [19] is a Natural Language Processing library implemented in Python and Cython, supporting over 70 languages. It provides pretrained pipelines for tasks such as tagging, parsing, named entity recognition, and text classification. HuSpaCy [20] is the Hungarian adaptation of SpaCy, including pretrained embedding models. We employed the 300-dimensional Hungarian CBOW embeddings[4] provided by HuSpaCy, trained on the Hungarian Webcorpus 2.0.

D. ELMo

ELMo (Embeddings from Language Models) [21] generates contextualized word embeddings using a bidirectional long short-term memory (LSTM) language model. Unlike static embeddings, ELMo captures polysemy and context-dependent meanings, improving performance across multiple NLP tasks. We used the Hungarian ELMo embeddings[5] provided by HIT-SCIR, trained on a Hungarian Wikipedia dump and the Hungarian portion of the Common Crawl.

E. BERT-based models

BERT (Bidirectional Encoder Representations from Transformers) [7] is a transformer-based model that, like ELMo, generates contextualized word embeddings. It is trained using a masked language model (MLM) objective, where the goal is to predict masked words within a sentence, and a next sentence prediction (NSP) objective, where the aim is to determine whether two sentences are consecutive. As a base model, we used huBERT [22], trained on the Hungarian Webcorpus 2.0. To have a baseline, we also included XLM-RoBERTa (later referred to as *XLM-R*) [23] with the same strategies.

For extracting static embeddings from BERT-based models, we employed the three methods mentioned in *Related Work*. *Decontextualized*, where only the word is inputted to the model, *Aggregate*, which pools embeddings of different contexts the word appears in, and *X2Static*, which uses a CBOW-inspired static word-embedding approach, leveraging contextual information from a teacher model to generate static embeddings. For both the *Aggregate* and *X2Static* methods, we used the same training data, which was a text file, containing

[4]https://huggingface.co/huspacy/hu_vectors_web_lg
[5]https://github.com/HIT-SCIR/ELMoForManyLangs

one sentence in a line, collected from Hungarian Wikipedia and the Hungarian Webcorpus[6].

Table I provides an overview of the models used in the study, including their dimensions and original vocabulary sizes.

TABLE I
MODEL SPECIFICATIONS

Model	Dimensions	Original Vocabulary
FastText	300	2,000,000
huBERT	768	-
EFNILEX	600	1,896,111
HuSpaCy	300	1,524,582
XLM-R	768	-
ELMo	1024	-

IV. DATASETS

A. Intrinsic Evaluation Datasets

Intrinsic evaluations assess the quality of embeddings independently, rather than through downstream tasks. They typically measure word similarity or evaluate embeddings on word analogy tasks.

Since no Hungarian word similarity datasets are available, we used the word analogy dataset developed by Makrai [24] for intrinsic evaluation. Inspired by the widely used analogy task introduced by Mikolov et al. [25], this dataset consists of structured word analogy questions in the form of *A:B :: C:D*, where the relationship between words A and B is expected to be analogous to the relationship between words C and D.

The dataset encompasses multiple linguistic and semantic relations, including morphological (e.g., singular-plural forms, verb conjugations), semantic (e.g., country-capital pairs), and syntactic analogies (e.g., grammatical roles).

To make comparisons fair, we restricted the embeddings to the vocabulary of their intersection, arriving at a vocabulary of 256 808 words. Table II provides a comparison of number of questions by category as they were originally, opposed to after the restrictions. There were three categories, where the ratio of the restricted and original questions fell below 90%: capital-world, family, and gram5-present-participle. For example, the capital-world category contained all the country-capital pairs. As smaller countries capital were not present in the vocabulary, the number of questions in this category significantly reduced.

TABLE II
COMPARISON OF ORIGINAL AND RESTRICTED NUMBER OF QUESTIONS BY CATEGORIES

Category	Original	Restricted	Ratio (%)
capital-common-countries	190	190	100.0%
capital-world	13695	5995	43.8%
county-ceter	171	171	100.0%
currency	435	406	93.3%
family	190	136	71.6%
gram1-adjective-to-adverb	780	780	100.0%
gram2-opposite	435	435	100.0%
gram3-comparative	780	780	100.0%
gram4-superlative	780	780	100.0%
gram5-present-participle	780	496	63.6%
gram6-nationality-adjective	820	741	90.4%
gram7-past-tense	780	780	100.0%
gram8-plural-noun	780	780	100.0%
gram9-plural-verb	780	780	100.0%

B. Extrinsic Evaluation Datasets

For extrinsic evaluation, we used NYTK-NerKor [26], a Hungarian gold-standard named entity-annotated corpus containing 1 million tokens. This dataset also includes POS tags, making it suitable for both Named Entity Recognition (NER) and Part-of-Speech (POS) tagging tasks. The corpus is divided into training, development, and test sets using an approximately 80-10-10 split. It comprises texts from diverse genres, including fiction, legal documents, news, web sources, and Wikipedia.

The fiction subcorpus includes novels from the Hungarian Electronic Library (MEK) and Project Gutenberg, as well as subtitles from OpenSubtitles. Legal texts originate from sources such as the EU Constitution, documents from the European Economic and Social Committee, DGT-Acquis, and JRC-Acquis. The news subcorpus draws from the European Commission's Press Release Database, Global Voices, and the NewsCrawl Corpus. Web texts are sourced from the Hungarian Webcorpus 2.0.

Named entity annotations are divided into four categories: Person, Location, Organization, and Miscellaneous. Although an updated version of the corpus [27] introduces approximately 30 entity types, we used the original version, as these four categories are most commonly utilized in NER tasks.

V. EXPERIMENTS

A. Intrinsic Evaluation

To assess performance on the analogy task, we employed two evaluation metrics. The first metric was overall accuracy, defined as the proportion of correctly answered questions, where the most likely prediction aligns with the correct answer. The second metric was the Mean Reciprocal Rank (MRR), a widely adopted evaluation measure that rewards not only the best prediction but also cases where the correct answer appears among the top-ranked candidates. Given a set of queries, MRR is computed as the average of the reciprocal ranks of the first relevant result for each query:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

where $|Q|$ denotes the total number of queries, and rank_i represents the rank position of the first relevant document for the ith query. To ensure computational feasibility, we restricted analysis to the top 10 ranked answers per query, assigning a score of 0 when the correct answer was absent. Given that different categories contained varying numbers of questions, we also report the average MRR across categories along with the overall MRR (which is the weighted sum).

For static embedding extraction from BERT-based models, we utilized three previously described methods: *decontextualized*, *aggregate*, and *X2Static*. These are indicated by subscripts in the model names. For instance, $\text{huBERT}_{\text{de}}$ refers

[6]http://mokk.bme.hu/resources/webcorpus/

to static embeddings extracted using the *decontextualized* method, while huBERT$_{agg}$ and huBERT$_{x2}$ correspond to the *aggregate* and *X2Static* methods, respectively.

Table III presents the results of the analogy tasks. FastText exhibited superior performance, achieving an overall accuracy of 71% and an MRR score of 0.77. The second-best model, huBERT, attained 49% accuracy and an MRR of 0.58. The remaining models formed two distinct groups based on overall accuracy. The first group, comprising EFNILEX, HuSpaCy, and XLM-R$_{x2}$, yielded an MRR of approximately 0.46 and an accuracy of around 38%. The second group, including huBERT$_{agg}$, XLM-R$_{agg}$, ELMo, and huBERT$_{de}$, exhibited an MRR around 0.23 with an accuracy near 18%. XLM-R$_{de}$ performed the worst, with both metrics falling below 2%, indicating that the *decontextualized* embedding extraction was unsuitable for this task.

Analyzing overall accuracy, we observe that among the top four models are the three static embedding models, along with only one BERT-based model. This suggests that static embeddings remain competitive for intrinsic tasks despite their lack of contextual adaptability compared to transformer-based models. Notably, huBERT$_{x2}$ performed comparably to static embeddings, indicating that the *X2Static* method is a promising technique for deriving static representations from BERT-based models.

The average MRR scores follow a similar pattern to overall accuracy but reveal a more nuanced ranking, as it considers every category equally important, regardless of its size. While the top five models remain consistent, the performance gaps narrow. This discrepancy is likely due to category imbalances, where some categories contain significantly more questions than others. Among the lower-performing models, ordering shifts: ELMo surpasses both *aggregate* models, while huBERT$_{de}$ outperforms huBERT$_{agg}$.

Examining the performance of the three embedding extraction methods in terms of overall accuracy, a clear trend emerges. The *decontextualized* method consistently underperforms in both huBERT and XLM-R. The *aggregate* method offers a marginal improvement over *decontextualized* in huBERT, yet it demonstrates a substantial boost in XLM-R. This improvement may stem from XLM-R's multilingual nature, where isolated Hungarian words may provide insufficient input, whereas contextualized sentences enable better representations. The *X2Static* method outperforms both alternatives across both models, approaching the effectiveness of dedicated static embeddings. However, when considering average MRR scores, while *X2Static* remains the best performer, huBERT$_{de}$ surpasses huBERT$_{agg}$.

Category-level results reveal substantial variation in model performance. FastText dominates, achieving the highest scores in 7 out of 14 categories, with the largest margins observed in *capital-common-countries*, *capital-world*, and *gram2-opposite*. HuBERT prevails in three categories (*family*, *gram9-plural-verb*, and *gram7-past-tense*), drawing with EFNILEX in the latter. EFNILEX also leads in *gram4-superlative*. HuSpaCy stands out in *gram5-present-participle*, significantly outperforming other models. Notably, this category experienced the most substantial question reduction (37%) due to vocabulary constraints. HuSpaCy achieved a 72% MRR score in this category, despite its average MRR of 0.56.

XLM-R$_{x2}$ demonstrates strong performance, leading in the *gram1-adjective-to-adverb* category and slightly outperforming others in *gram3-comparative*. An interesting observation is that while ELMo and huBERT$_{de}$ exhibit similar overall performance, their category-level results diverge significantly, with each model excelling in different areas. In total, static embeddings win 10 of 14 categories, BERT-based models obtain three victories, and one category results in a tie. These findings underscore that category-level performance can vary considerably, making overall scores an incomplete representation of a models strengths.

When analyzing embedding extraction techniques for huBERT at the category level, *X2Static* consistently outperforms the other methods in all but one category (*gram8-plural-noun*), where the *decontextualized* method proves to be superior. Comparing *aggregate* and *decontextualized* approaches, results are less conclusive. For huBERT, *aggregate* excels in only four categories, while *decontextualized* leads in nine, with one tie. XLM-R follows a more distinct hierarchy, with *X2Static* outperforming all other methods, *aggregate* ranking second, and *decontextualized* consistently underperforming.

These results collectively highlight the effectiveness of static embeddings for analogy tasks, the viability of the *X2Static* approach for extracting static embeddings from transformer-based models, and the varying impact of embedding extraction techniques across different linguistic categories.

B. Extrinsic Evaluation

For the extrinsic evaluation, we employed a single-layer bidirectional LSTM with a dropout rate of 0.5. The bidirectional LSTM was chosen for its ability to capture contextual information from both directions, which is particularly beneficial for tasks such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging. To assess the performance of the embeddings, we experimented with varying hidden sizes (1, 2, 4, 8, 16, 32, and 64). While increasing the hidden size generally enhances model performance, the objective was to evaluate the embeddings themselves, including their behavior under constrained settings.

To handle out-of-vocabulary words in the training data, we represented them using vectors sampled from a normal distribution with a mean of 0 and a standard deviation of 0.6 ($\mathcal{N}(0, 0.6)$). The model architecture concluded with a softmax activation function. Training was conducted for five epochs with a batch size of 32, utilizing the Adam optimizer with its default parameter settings. Categorical cross-entropy was employed as the loss function.

Table IV summarizes the performance of the NER models on the test set. As the primary objective was to assess the quality of the embeddings rather than to achieve state-of-the-art results, accuracy was selected as the evaluation metric. While suitable for the purposes of our comparative analysis, it should be noted that this metric is not directly comparable to those reported by leading models in the field, which are typically evaluated using the F1 score [28].

TABLE III
COMPARISON OF PERFORMANCE ACROSS MODELS (MRR)

Category	FastText	huBERT$_{x2}$	EFNILEX	HuSpaCy	XLM-R$_{x2}$	huBERT$_{agg}$	XLM-R$_{agg}$	ELMo	huBERT$_{de}$	XLM-R$_{de}$
capital-common-countries	**0.77**	0.58	0.45	0.44	0.40	0.25	0.26	0.09	0.17	0.01
capital-world	**0.83**	0.50	0.28	0.25	0.30	0.23	0.17	0.03	0.08	0.00
county-center	**0.88**	0.76	0.31	0.47	0.25	0.18	0.07	0.09	0.24	0.00
currency	**0.31**	0.10	0.19	0.15	0.09	0.07	0.09	0.12	0.07	0.00
family	0.66	**0.67**	0.40	0.59	0.46	0.22	0.25	0.33	0.30	0.05
gram1-adjective-to-adverb	0.63	0.59	0.37	0.61	**0.78**	0.20	0.26	0.13	0.25	0.07
gram2-opposite	**0.43**	0.16	0.29	0.24	0.17	0.01	0.04	0.10	0.07	0.01
gram3-comparative	0.76	0.81	0.75	0.74	**0.82**	0.30	0.39	0.47	0.38	0.10
gram4-superlative	0.68	0.63	**0.72**	0.59	0.29	0.21	0.19	0.27	0.22	0.00
gram5-present-participle	0.55	0.17	0.11	**0.72**	0.12	0.02	0.02	0.01	0.15	0.00
gram6-nationality-adjective	**0.91**	0.87	0.68	0.61	0.64	0.37	0.37	0.09	0.22	0.01
gram7-past-tense	0.82	**0.95**	**0.95**	0.86	0.91	0.18	0.17	0.82	0.49	0.01
gram8-plural-noun	**0.77**	0.61	0.66	0.68	0.66	0.28	0.34	0.67	0.70	0.06
gram9-plural-verb	0.94	**0.97**	0.95	0.87	0.87	0.67	0.68	0.86	0.55	0.02
Average MRR	**0.71**	0.60	0.51	0.56	0.48	0.23	0.24	0.29	0.28	0.02
Overall MRR	**0.77**	0.58	0.46	0.46	0.45	0.24	0.23	0.22	0.22	0.02
Overall accuracy	**0.71**	0.49	0.39	0.38	0.37	0.18	0.18	0.18	0.17	0.01

TABLE IV
MODEL PERFORMANCE ACROSS HIDDEN SIZES FOR NER (%)

Model Name	Hidden Size						
	1	2	4	8	16	32	64
ELMo	94.81	**96.11**	**96.44**	**97.01**	**97.39**	**97.54**	**97.62**
huBERT$_{x2}$	93.55	95.42	96.02	96.68	97.24	97.40	97.49
RoBERTa$_{x2}$	94.79	95.48	96.16	96.55	97.20	97.33	97.44
huBERT$_{de}$	**94.89**	94.12	96.22	96.76	97.16	97.24	97.38
HuSpaCy	94.16	94.71	95.84	96.23	96.54	96.92	97.13
huBERT$_{agg}$	94.92	94.97	95.16	95.95	96.61	96.94	97.01
EFNILEX	94.48	94.86	95.49	96.09	96.46	96.71	96.80
FastText	94.59	94.89	95.62	96.32	96.56	96.68	96.78
XLM-R$_{agg}$	85.39	94.67	94.37	95.07	95.12	95.30	95.68
XLM-R$_{de}$	94.79	88.59	94.97	95.03	95.10	95.22	95.37

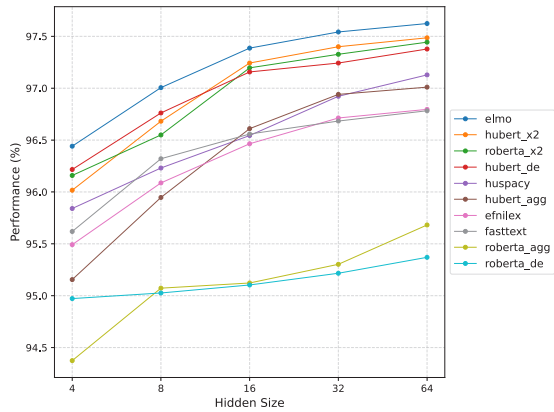


Fig. 1: Model performance across hidden sizes for NER (%)

As shown in Table IV, most models demonstrate a consistent improvement in performance with increasing hidden size. However, exceptions to this trend include XLM-R$_{agg}$ (between hidden sizes 2 and 4), as well as huBERT$_{de}$ and XLM-R$_{de}$ (between sizes 1 and 2), with the latter experiencing a particularly sharp decline. To complement the tabular data, Figure 1 offers a visual summary that aids intuitive comparison, albeit with a slight trade-off in numerical precision. For clarity, hidden sizes 1 and 2 were excluded from the plot due to their disproportionate effect on the plot scale, which hindered the visibility of differences among the larger hidden sizes.

Regarding overall performance, XLM-R$_{agg}$ and XLM-R$_{de}$ were the lowest-performing models, indicating their relative inefficacy for this task. In contrast—and somewhat unexpectedly—ELMo emerges as the top-performing model across all hidden sizes of two and above, despite its weaker results in the analogy task. It may come from the fact, that ELMo is originally trained using a bidirectional LSTM, similarly to the model used in the extrinsic evaluations. For larger hidden sizes, it is followed by the two *X2Static* models, huBERT$_{x2}$ and XLM-R$_{x2}$, which maintain close performance levels but exchange rankings depending on the hidden size. Notably, at smaller hidden sizes, huBERT$_{de}$ outperforms both *X2Static* models. The remaining four models—HuSpaCy, huBERT$_{agg}$, EFNILEX, and FastText—consistently occupy the middle range, positioned between the top four and the bottom two, with their relative rankings fluctuating based on hidden size.

In terms of extraction methodologies, the *X2Static* approach remains the most effective, consistent with the intrinsic task results. HuBERT performs better in its *decontextualized* configuration compared to its *aggregate* variant in most cases, likely due to the aggregate model incorporating the full contextual embedding rather than isolating word-level information. XLM-R exhibits an opposite trend, albeit with a less pronounced difference than in the analogy task.

These findings suggest that the top-performing four models are all derived from dynamic embeddings, highlighting their advantage for this task. Nevertheless, purely static embeddings are not far behind in performance, demonstrating competitiveness. Notably, huBERT$_{agg}$ is the only BERT-based model to be consistently outperformed—depending on the hidden size—by static embedding models.

Table V summarizes the numerical results for the POS tagging task, while Figure 2 offers a visual depiction of model performance. As with Figure 1, hidden size 1 is excluded to preserve visual clarity and enable more meaningful comparison across the remaining configurations.

Among the evaluated models, XLM-R$_{agg}$ and XLM-R$_{de}$ exhibited the lowest performance, while huBERT$_{agg}$ outperformed them but remained behind the rest. Consistently,

TABLE V
PERFORMANCE OF VARIOUS MODELS ACROSS DIFFERENT HIDDEN SIZES IN POS TAGGING (%)

Model Name	Hidden Size						
	1	2	4	8	16	32	64
ELMo	45.69	**83.94**	**88.30**	90.56	**93.00**	**94.00**	**94.58**
huBERT$_{x2}$	**60.40**	73.06	84.59	**91.34**	92.50	93.38	93.95
RoBERTa$_{x2}$	40.89	72.65	87.38	90.21	92.36	93.46	93.76
HuSpaCy	51.20	69.16	83.09	89.31	91.26	92.47	93.33
huBERT$_{de}$	15.53	72.69	86.80	88.86	91.05	92.98	93.17
EFNILEX	38.15	67.11	78.81	85.92	87.91	89.38	90.48
FastText	46.09	69.46	80.01	86.09	88.33	89.47	89.99
huBERT$_{agg}$	25.87	48.99	65.64	76.76	82.40	85.60	87.39
XLM-R$_{de}$	32.19	45.43	54.20	60.57	67.95	71.20	78.37
XLM-R$_{agg}$	15.63	35.41	39.23	49.14	55.57	53.53	52.57

ELMo achieved the highest accuracy across most hidden sizes, followed closely by the two *X2Static* models, huBERT$_{x2}$ and XLM-R$_{x2}$.

The remaining models can be grouped into two performance tiers. HuBERT$_{de}$ and HuSpaCy formed the stronger pair, demonstrating similar results. EFNILEX and FastText constituted the second tier, with FastText maintaining a slight edge in configurations with less than 64 hidden sizes.

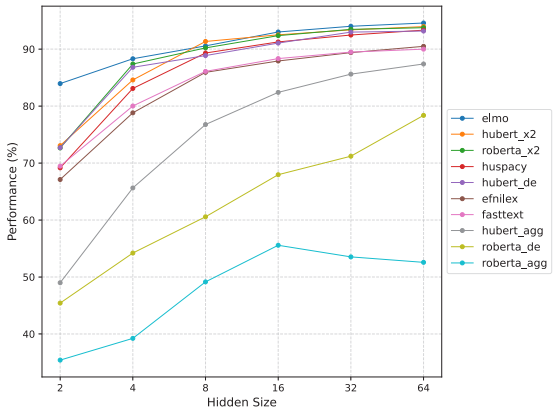


Fig. 2: Performance of various models across different hidden sizes in POS tagging (%)

In alignment with findings from other experiments, the *X2Static* method emerged as the best-performing approach. Notably, both huBERT and XLM-R achieved superior results when employing the *decontextualized* method rather than the *aggregate* one.

As observed in the Named Entity Recognition (NER) task, static embeddings derived from dynamic models consistently outperformed purely static models. Despite this trend, the gap in performance remained relatively small, reinforcing the effectiveness of static embeddings in POS tagging tasks.

VI. CONCLUSION AND FUTURE WORK

In this paper, we conducted a comprehensive analysis of various static word embeddings for Hungarian, alongside static embeddings derived from BERT-based models, evaluating them on both intrinsic and extrinsic tasks. To ensure a fair comparison, all models were restricted to a common vocabulary.

For intrinsic evaluation, we employed an analogy task, where FastText demonstrated superior performance, achieving an overall accuracy of 71% and a Mean Reciprocal Rank (MRR) score of 0.77. Among the BERT-based models, huBERT$_{x2}$ emerged as the best performer, with an accuracy of 49% and an MRR score of 0.58. Notably, the *X2Static* method for extracting static embeddings from BERT-based models outperformed both the *decontextualized* and *aggregate* methods, even rivaling traditional static embeddings in intrinsic evaluations.

For extrinsic evaluation, we utilized a single-layer bidirectional LSTM with varying hidden sizes to assess the effectiveness of the embeddings in downstream tasks. The ELMo embeddings consistently outperformed other models in both Named Entity Recognition (NER) and Part-of-Speech (POS) tagging tasks. The *X2Static* method remained the most effective for extracting static embeddings from BERT-based models, while static embeddings derived from dynamic models outperformed purely static models in both tasks.

This piece of research paves the way for multiple avenues of future exploration. A key direction would be the development of a new intrinsic evaluation dataset for Hungarian, as the analogy question dataset remains the sole benchmark for such evaluations. Additionally, investigating the impact of dimensionality reduction on model performance could yield insights into the trade-offs between efficiency and accuracy, as the models analyzed in this study were trained with varying dimensional settings. Furthermore, the methodology employed here could be extended to evaluate other Hungarian BERT-based models, broadening the scope of comparative analyses. Another significant area of investigation is the exploration of Hungarian GPT-based models ([29], [30]), which were not included in this study due to the lack of a clear methodology for extracting static embeddings from such architectures.

In the spirit of open science, the training scripts, evaluation codes, restricted vocabulary, and the extracted huBERT$_{x2}$ embeddings are made publicly available, facilitating further research and reproducibility in the field.

REFERENCES

[1] Yoshua Bengio et al. "A neural probabilistic language model". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435. DOI: 10.5555/944919.944966.

[2] Ronan Collobert and Jason Weston. "A unified architecture for natural language processing: deep neural networks with multitask learning". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 160–167. ISBN: 9781605582054. DOI: 10.1145/1390156.1390177. URL: https://doi.org/10.1145/1390156.1390177.

[3] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. DOI: 10.48550/arXiv.1301.3781. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781

[4] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/.

[5] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics 5* (July 2016). DOI: 10.1162/tacl_a_00051.

[6] Ashish Vaswani et al. "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964. DOI: 10.5555/3295222.3295349.

[7] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[8] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[9] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Oct. 2019. DOI: 10.48550/arXiv.1910.10683.

[10] Rishi Bommasani, Kelly Davis, and Claire Cardie. "Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 4758–4781. DOI: 10.18653/v1/2020.acl-main.431. URL: https://aclanthology.org/2020.acl-main.431/.

[11] Prakhar Gupta and Martin Jaggi. "Obtaining Better Static Word Embeddings Using Contextual Embedding Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 5241–5253. DOI: 10.18653/v1/2021.acl-long.408. URL: https://aclanthology.org/2021.acl-long.408/.

[12] Karahan Sarıtaş, Cahid Arda Öz, and Tunga Güngör. "A comprehensive analysis of static word embeddings for Turkish". In: *Expert Systems with Applications 252* (2024). ISSN: 0957-4174. DOI: 10.1016/j.eswa.2024.124123. URL: http://dx.doi.org/10.1016/j.eswa.2024.124123.

[13] Bálint Döbrössy et al. "Investigating Sub-Word Embedding Strategies for the Morphologically Rich and Free Phrase-Order Hungarian". In: *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Ed. by Isabelle Augenstein et al. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 187–193. DOI: 10.18653/v1/W19-4321. URL: https://aclanthology.org/W19-4321/.

[14] Judit Ács et al. *Evaluating Contextualized Language Models for Hungarian*. Feb. 2021. DOI: 10.48550/arXiv.2102.10848.

[15] Judit Ács, Ákos Kádár, and Andras Kornai. "Subword Pooling Makes a Difference". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics*, Apr. 2021, pp. 2284–2295. DOI: 10.18653/v1/2021.eacl-main.194. URL: https://aclanthology.org/2021.eacl-main.194/.

[16] Márton Makrai. "Filtering Wiktionary Triangles by Linear Mbetween Distributed Word Models". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2766–2770. URL: https://aclanthology.org/L16-1439/.

[17] Péter Halácsy et al. "Creating Open Language Resources for Hungarian". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Ed. by Maria Teresa Lino et al. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. URL: https://aclanthology.org/L04-1320/.

[18] Tamás Váradi. "The Hungarian National Corpus". In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Ed. by Manuel González Rodríguez and Carmen Paz Suarez Araujo. Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), May 2002. URL: https://aclanthology.org/L02-1217/.

[19] Matthew Honnibal et al. "spaCy: Industrial-strength Natural Language Processing in Python". In: (2020). DOI: 10.5281/zenodo.1212303.

[20] György "Orosz et al. ""Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines"". In: *"Text, Speech, and Dialogue"*. Ed. by Kamil "Ekštein, František Pártl, and Miloslav" Konopík. "Cham": "Springer Nature Switzerland", 2023, pp. 58–69. ISBN: 978-3-031-40498-6. DOI: 10.1007/978-3-031-40498-6_6.

[21] Matthew E. Peters et al. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202/.

[22] Dávid Márk Nemeskey. "Introducing huBERT". In: *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*. Szeged, 2021, TBA.

[23] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747/.

[24] Márton Makrai. "Comparison of distributed language models on medium-resourced languages". In: *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*. Ed. by Attila Tanács, Viktor Varga, and Veronika Vincze. 2015. ISBN: 978-963-306-359-0.

[25] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. DOI: 10.48550/arXiv.1301.3781. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

[26] Eszter Simon and Noémi Vadász. "Introducing NYTKNerKor, A Gold Standard Hungarian Named Entity Annotated Corpus". In: *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*. Ed. by Kamil Ekstein, Frantisek Pártl, and Miloslav Konopík. Vol. 12848. Lecture Notes in Computer Science. Springer, 2021, pp. 222–234. DOI: 10.1007/978-3-030-83527-9_19.

[27] Attila Novák and Borbála Novák. "NerKor 1.41e". In: *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. 2022.

[28] Simon Eszter et al. "Az NYTK-NerKor több szempontú kiértékelése". In: *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, 2022, pp. 403–416.

[29] Zijian Győző Yang et al. "Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre". In: *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*. Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Intézet, 2023, pp. 247–262.

[30] Kristóf Szentmihályi et al. "Pretraining GPT-style models in Hungarian". In: *Infocommunications Journal* (2025).

Máté Gedeon earned his bachelor's degree in Mathematics from the Budapest University of Technology in 2023. He is currently pursuing an MSc in Mathematics at the same institution, specializing in data science. His primary research interests lie in the fields of Natural Language Processing (NLP) and Automatic Speech Recognition (ASR).

IP Packet Forwarding Performance Comparison of the FD.io VPP and the Linux Kernel

Melinda Kosák, and Gábor Lencse

***Abstract*—There are numerous free software solutions for IPv4 or IPv6 packet forwarding. The Fast Data Project / Vector Packet Processing (FD.io VPP) is a novel and prominent solution. This paper investigates its performance and scalability compared to that of the Linux kernel. The investigation was conducted in accordance with the requirements outlined in the relevant Request for Comments (RFC) documents (RFC 2544, RFC 4814, and RFC 5180) using the siitperf measurement software. Two different test environments were used to eliminate the potential hardware-specific side effects and to gain insight into the performance and scalability of the IPv4 and IPv6 packet forwarding capability of the two investigated solutions. It was found that FD.io VPP outperformed the Linux kernel by approximately an order of magnitude. The configuration of FD.io VPP, along with the details of the measurements, are provided, and the results are presented and analyzed in the paper.**

***Index Terms*—FD.io VPP, IPv4, IPv6, Linux kernel, packet forwarding**

I. INTRODUCTION

THE Internet is an essential part of our life. All data of our daily communication is carried in packets by the Internet Protocol (IP). At the time of writing, there are two versions of IP: the older IPv4 version and the newer, increasingly more adopted IPv6. For IPv4 and IPv6 packet forwarding, there are numerous free software solutions, such as the Linux kernel, and one with exceptionally high performance called FD.io VPP. This article focuses on FD.io VPP, as its developers claim: "The Fast Data Project (FD.io) is an open-source project aimed at providing the world's fastest and most secure networking data plane through Vector Packet Processing (VPP)." [1]. We examine this proposition from a performance perspective. In this paper, we compare the IPv4 and IPv6 packet forwarding performance and scalability of the Linux kernel and FD.io VPP. By scalability, we refer to how their performance increases with the number of CPU cores utilized. First, we complete the performance and scalability test of the Linux kernel to establish a basis for comparison. Then we continue with the examination of the performance and scalability of FD.io VPP. For our tests, we use two different hardware environments.

Submittted November 18, 2024.

M. Kosák is with the Cybersecurity and Network Technologies Research Group of the Faculty of Mechanical Engineering, Informatics, and Electrical Engineering, Széchenyi István University, Győr, Hungary (e-mail: kosakmeli@gmail.com).

G. Lencse is with the Department of Telecommunications, Széchenyi István University, Győr, Hungary (e-mail: lencse@sze.hu).

DOI: 10.36244/ICJ.2025.2.5

The remainder of this paper is organized as follows. In section II, a brief introduction is given to the theoretical background of the performance measurements of network interconnect devices on the basis of the relevant Internet Engineering Task Force (IETF) Request for Comments (RFC) documents. Section III presents a short survey of related works. Section IV is an overview of the software used for our measurements. Section V discloses the relevant details of our measurement environments. In section VI, our measurement results are presented and analyzed. Section VII covers the discussion and our plans for future research. Section VIII concludes our paper.

II. THEORETICAL BACKGROUND

A. RFC 2544

The main purpose of the RFC 2544 [2] is to define how to measure the performance of network interconnect devices in an objective and repeatable way. The most important aspects include measurement setup, Device Under Test (DUT) setup, frame format and frames sizes, and testing duration.

The RFC lists three different measurement setups of which we used the first one.

The RFC recommends the usage of the following Ethernet frame sizes: 64, 128, 256, 512, 1024, 1280, 1518. It defines TCP/IP over Ethernet frame formats. They are: learning frame, routing update frame, management query frame and test frame. The test frame is used for different benchmarking tests like throughput, latency, frame loss rate, back-to-back frames, system recovery, and reset. From among them, throughput is the most essential one.

To measure the performance of routers, an IP address range was reserved, 198.18.0.0/15. The lower half of the range (198.18.0.0/16) was used on the left side of the configuration shown in Figure 1, and the upper half (198.19.0.0/16) on the right side. The interfaces of the DUT were assigned the IP address ending in 1 of the domains (198.18.0.1 and 198.19.0.1).

It should be noted that RFC 2544 requires testing with bidirectional traffic and it was used in all our measurements.

B. RFC 4814

RFC 4814 [3] covers several different topics. During our measurements, we used the pseudorandom port numbers, which are recommended in section 4.5 of the aforementioned RFC. The network cards in use today are capable of distributing interrupts caused by the packet arrivals to different CPU cores for processing. This is done using a hash function that takes the source and destination IP addresses, as well as the TCP or UDP

source and destination port numbers in incoming packets, as input parameters. This process is called Receive-Side Scaling (RSS) [4], which aims to achieve scalability. To support RSS, RFC 4814 recommends the usage of pseudorandom port numbers. The recommended ranges are:

- for source port numbers: 1024 – 65535
- for destination port numbers: 1 – 49151

C. RFC 5180

RFC 5180 [5] is highly similar to RFC 2544 in several ways, but while RFC 2544 focused on IPv4, RFC 5180 deals with IPv6. Similar to IPv4, an IPv6 address space was reserved for benchmarking. The reserved range is 2001:2::/48. As in RFC 2544, the address range should be halved.

III. RELATED WORK

As for peer-reviewed research papers about measuring IPv4 and IPv6 packet forwarding performance of FD.io VPP according to the current industry standards laid down by the RFCs mentioned above, we found only our own conference paper [6], which is extended in our current paper with further measurements.

However, several other recent research papers recommended FD.io VPP for IPv4 or IPv6 packet forwarding.

For example, Slavic and Krajnovic [7] proposed that open-source software and commodity hardware may replace the router vendors' products. They recommended the usage of FD.io VPP and some further software. However, they did not present any benchmarking measurements that were RFC 2544 or RFC 5180 compliant.

FD.io VPP was used as the packet forwarding solution of a network emulator called “CNNet” [8].

FD.io VPP is integrated with a custom control plane to provide a high-performance, low-cost software cloud gateway for accelerating virtual cloud networks [9]. Although certain performance data are included in the paper, the testing conditions are not mentioned. Their measurements were not RFC 2544 compliant because of the traffic generator used.

Another paper gives an important insight into the issue of why user-space solutions may outperform the interrupt-based ones [10].

IV. SOFTWARE USED FOR MEASUREMENTS

In this section, a brief summary is provided on DPDK, the Linux kernel, Non-Uniform Memory Access (NUMA), FD.io VPP, and siitperf.

A. DPDK

The Data Plane Development Kit (DPDK) [11] is an open-source software, with Linux based user platform, that was designed to improve packet processing speeds. DPDK enables the rapid development of high-speed data packet networking applications. DPDK achieves fast packet processing by consisting of libraries and drivers that bypass the operating system’s network stack. DPDK-based programs can send and receive approximately an order of magnitude more packets per CPU core than those using the Linux kernel [11].

B. Linux kernel

During packet forwarding, the Linux kernel processes incoming network packets and forwards them to the appropriate destination. In the Linux kernel, both scalar packet processing and RSS play an important role in packet delivery. During the scalar packet processing, the kernel individually processes the incoming packets. Just one packet is taken by an interrupt function (by default) from the network interface, then it works through a series of functions [1]. This method is simple, but its efficiency can be limited as it requires the same call chain to be executed for each packet. This can be time-consuming and place a strain on the processor and caches. With RSS, the kernel can handle heavy loads more efficiently and distribute packet processing across multiple CPUs in the system, increasing performance and reducing latency.

C. Non-Uniform Memory Access

NUMA is a multiprocessor system design where memory access time depends on the position of memory relative to the processor: each processor accesses its own local memory faster than the local memory of another processor [12]. Whereas NUMA is necessary to support scalability, our results show the consequences of its usage when a CPU core belongs to a different NUMA node than the Network Interface Card (NIC) it communicates with.

D. FD.io VPP

The Fast Data Project (FD.io) [1] introduced Vector Packet Processing (VPP) that can handle high performance traffic. It can be used on multiple platforms. Vector packet processing can receive multiple packets at once and pass this group, known as a packet vector, to the processing function, which then processes it, thereby saving time. The Packet Processing Graph (PPG) is at the heart of the FD.io VPP design. FD.io VPP collects a vector of packets from the RX rings, up to 256 packets in a single vector. The received packets are then traversed through the nodes of the PPG in the vector, with each graph node representing network processing that is applied to each packet. FD.io VPP can be used with or without DPDK. We used it with DPDK.

E. Siitperf

Siitperf [13] was running on our Tester server. The name of siitperf comes from the fact that it was originally designed to measure the performance of Stateless IP/ICMP Translation (SIIT) gateways. Due to its flexibility, it is also suitable for measuring the performance of IPv4 and IPv6 packet forwarders (routers). Siitperf uses DPDK to achieve a sufficiently high performance [13]. It should be noted that siitperf reports the results as packets per second *per direction*. When bidirectional traffic is used, the number of all frames forwarded is double the value reported.

As siitperf supports the throughput, latency, frame loss rate and packet delay variation measurement procedures of RFC 8219 [14], and the throughput measurement procedure of RFC 2544 was incorporated in RFC 5180 and then in RFC 8219 without any changes, siitperf could be used for measuring IPv4 and IPv6 throughput according to the requirements of RFC 2544 and RFC 5180, respectively.

Siitperf is a collection of binaries and bash shell scripts. The binaries execute an elementary step of the given measurement procedure. For example, the siitperf-tp binary performs a 60-second long throughput test, and the binary-rate-alg.sh script performs the binary search by calling the siitperf-tp binary and providing the appropriate parameter values as command line arguments. In contrast, other parameters that do not change during the consecutive steps of the binary search are read from the siitperf.conf configuration file.

As required by the throughput measurement procedures of RFC 2544 or RFC 5180, siitperf-tp sends bidirectional IPv4 or IPv6 traffic and counts the received frames. Based on the reported values, the shell script determines throughput, which is the highest frame rate at which the DUT can forward all test frames without loss.

Further details of the design, implementation, operation and performance of siitperf can be found in several research papers. The original design, which relied on fixed port numbers, was disclosed in [13]. The extension to support RFC 4814 pseudorandom port numbers was documented in [15]. The accuracy of siitperf was checked by comparing the IPv4 throughput of the same DUT determined by siitperf and the RFC 2544-compliant commercial Anritsu MP1590B Network Performance Tester [16]. The theory and practice of extending siitperf for stateful tests was published in [17]. Finally, support for pseudorandom IP addresses was added, as described in [18].

V. TEST ENVIRONMENT

A. The Structure of the Test Network

Two test systems were used. The first one consisted of Dell PowerEdge R620 servers. Each had two 6-core Intel Xeon E5-2620 processors and 32 GB 1600 MHz DDR3 SDRAM configured as 1333 MT/s. The second one contained Dell PowerEdge R730 servers. Each had two 8-core Intel Xeon E5-2667 v4 CPUs and 128 GB 2666 MHz DDR4 SDRAM configured as 2400 MT/s. We installed an Intel X540 10G/1G network interface card (NIC) in each of them, using the two 10 GbE ports for the measurements. The servers were directly connected with Cat6 UTP patch cables. The test setup is shown in Fig. 1.

To achieve stable measurement results, we switched off hyper-threading and set the CPU clock frequency of the servers to a fixed rate at their nominal clock frequency using the tlp Linux package, namely to 2 GHz and to 3.2 GHz for the R620 and the R730 servers, respectively.

As for drivers for the 10GbE ports of the X540 NIC, ixgbe and uio_pci_generic were used with the Linux kernel and with FD.io.VPP.

It should be noted that all servers used had two NUMA nodes, where the 10GbE network interfaces and the CPU cores with even serial numbers (core 0, core 2, core 4, etc.) belonged to NUMA node 0 and the CPU cores with odd serial numbers (core 1, core 3, core 5, etc.) belonged to NUMA node 1.

B. Performance of the Tester

As we did not use a commercial network performance tester to perform the measurements but instead employed our own software tester called siitperf, which ran on the same type of servers as the DUT, it was important to avoid the situation that the Tester could become a bottleneck. To achieve this, we conducted a loopback test: the two interfaces of the Tester were interconnected by a direct cable, leaving out the DUT, and a throughput test was performed. (This was called the "self-test of the Tester" in our previous papers about siitperf.)

This test was only performed with the R620 Tester using IPv4. The result was highly stable: 6.03 Mfps. The test was not repeated with IPv6 traffic because, according to our experience, it would not make a significant difference. (Please refer to Table 4 and Table 5 of [18].) We did not need to perform the test with the R730 server because we knew from our previous experiments that the X540 NIC formed the bottleneck, as it can do about 7.1-7.2 Mfps. (As already mentioned, these rates were measured with bidirectional traffic and are to be understood as *per direction* rates.)

C. Configuration of FD.io VPP

We followed the installation guide from the official FD.io webpage [1]. We installed the following packages: libvppinfra, vpp, vpp-plugin-core, vpp-plugin-dpdk. During FD.io VPP measurements, we assigned our interfaces to DPDK with the uio_pci_generic driver. The configuration of FD.io VPP was done with the following commands:

```
set interface ip address \
TenGigabitEthernet1/0/0 198.18.0.1/24
set interface ip address \
TenGigabitEthernet1/0/1 198.19.0.1/24
set interface state TenGigabitEthernet1/0/0 up
set interface state TenGigabitEthernet1/0/1 up
set ip neighbor TenGigabitEthernet1/0/0 \
198.18.0.2 24:6e:96:3b:fb:00
set ip neighbor TenGigabitEthernet1/0/1 \
198.19.0.2 24:6e:96:3b:fb:02
```

The latter two commands were necessary because siitperf cannot respond to ARP requests. As a result, we had to set the ARP table entries manually.

The IPv6 configuration was similar, using IPv6 addresses instead of the IPv4 addresses. In that case we set the NDP table entries similar to the ARP table entries.

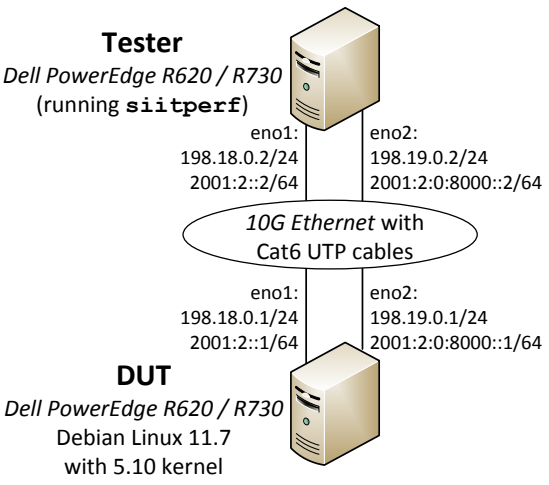


Fig. 1. Test setup for benchmarking FD.io VPP and the Linux kernel.

TABLE I
THROUGHPUT OF IPV4 PACKET FORWARDING OF THE LINUX KERNEL AS A FUNCTION OF THE NUMBER OF ACTIVE CPU CORES, E5-2620 @ 2 GHZ

Number of CPU cores	1	2	4	6	8	12
median (fps)	442,782	765,223	1,524,287	2,265,108	2,788,933	3,923,150
minimum (fps)	442,352	764,369	1,521,483	2,218,749	2,781,247	3,874,999
maximum (fps)	443,848	768,128	1,525,409	2,268,067	2,790,527	3,929,688
average (fps)	442,889	765,329	1,524,244	2,262,774	2,787,973	3,918,181
standard deviation	363.64	814.01	981.57	10,549.55	2,234.20	1,5163.08
dispersion (%)	0.34	0.49	0.26	2.18	0.33	1.39
relative to half as many cores	--	1.7282	1.9920	--	1.8297	1.7320
relative scale-up	1	0.8641	0.8606	0.8526	0.7873	0.7384

In VPP, for setting the number of CPU cores to be used, we had to modify the /etc/vpp/startup.conf configuration file. In the CPU segment, we changed the following settings:

```
main-core 2 # the main program runs on core 2
corelist-workers 4, 6 # 2 workers (on 4 and 6)
```

VI. MEASUREMENTS AND RESULTS

As for frame sizes, 64-byte and 84-byte test frames were used for IPv4 and IPv6 respectively. These are the smallest frame sizes allowed by siitperf. The rationale behind this choice was to make the CPU's processing capacity the bottleneck (limiting the maximum frame rate), rather than the packet transmission capacity of the network interface card.

It should be noted that since we used a general-purpose operating system, random events could occur during our measurements, potentially influencing the results. Therefore, each test was executed 20 times to achieve statistically reliable results.

As for summarizing function, both median and average was used. In the analysis, we primarily relied on median, because it is less sensitive to outliers than average.

To express the consistent or scattered nature of the results, we primarily relied on dispersion. It is defined by (1).

$$dispersion = \frac{maximum - minimim}{median} \cdot 100\% \qquad (1)$$

In addition, we also included the standard deviation.

As mentioned, in Section IV.E, siitperf reports the number of forwarded packets *per direction*. Therefore, our results should be interpreted accordingly (i.e., they should be multiplied by two to obtain the total number of forwarded packets per second).

A. Linux Kernel Packet Forwarding Performance of R620

The IPv4 and IPv6 packet forwarding performance of the Linux kernel was measured as a function of the number of active CPU cores. The number of active CPU cores was limited using the maxcpus=*n* kernel parameter, where *n* took the values of 1, 2, 4, 6, 8 and 12.

The IPv4 packet forwarding performance of the Linux kernel as a function of the number of active CPU cores is shown in Table I. At first glance, the results show that the performance of the DUT scaled up well with the increase of the number of active CPU cores. To facilitate a more detailed analysis of the results, the second-to-last line of the table shows the performance relative to having half as many active CPU cores, while the last line displays the relative scale-up, as defined by (2).

$$S(n) = P(n)/P(1)/n \qquad (2)$$

Where n is the number of active CPU cores, and P(n) is the performance measured (in frames per second) with n active CPU core. Its theoretical maximum value is 1.

A closer inspection of the results shows the following:

When using 2 CPU cores instead of 1 CPU core, the system performance did not double, but only increased by a factor of 1.7282. There are two main reasons for this phenomenon:

1. there is a performance cost of running a system with multiple cores compared to running a system with a single core,
2. CPU core 0 and the 10GbE Ethernet interfaces used for measurement belong to NUMA node 0, but CPU core 1 belongs to NUMA node 1, which means that CPU core 1 can only communicate with the network interface via core 0, and this communication overhead reduces the performance.

However, when we used 4 active CPU cores instead of 2, the performance doubled by a very good approximation (1.992 times). The explanation for this is very simple. The above factors were already present in the dual core system, so they did not cause any additional performance degradation in the quad core system.

TABLE II
THROUGHPUT OF IPV6 PACKET FORWARDING OF THE LINUX KERNEL AS A FUNCTION OF THE NUMBER OF ACTIVE CPU CORES, E5-2620 @ 2 GHZ

Number of CPU cores	1	2	4	6	8	12
median (fps)	425,782	732,299	1,466,585	2,176,369	2,658,915	3,766,926
minimum (fps)	424,951	726,561	1,374,999	1,999,999	2,617,186	3,761,709
maximum (fps)	426,026	733,467	1,466,974	2,178,284	2,660,414	3,772,095
average (fps)	425,628	731,888	1,461,745	2,156,176	2,656,771	3,766,972
standard deviation	291.61	1,481.93	20,459.50	54,645.21	9,443.81	2,679.58
dispersion (%)	0.25	0.94	6.27	8.19	1.63	0.28
relative to half as many cores	--	1.7199	2.0027	--	1.8130	1.7308
relative scale-up	1	0.8599	0.8611	0.8519	0.7806	0.7373

Further on, the performance of the 8-core system compared to a 4-core system is only 1.8297 times higher, and the comparison of performance of the 12-core system and the 6-core system shows only a 1.732 times increase. There may already be other reasons for this degraded increase, such as the fact that all CPU cores share the same network interfaces and, as their number increases, they are already interfering with each other to some extent by competing for the access to the network interfaces.

In conclusion, we observed that the performance of the system scaled up well: the throughput of the 12-core system instead the 1-core system increased from 442,782 fps to 3,923,150 fps, which means an 8.86 times increase.

The results of the throughput measurements for the IPv6 packet forwarding performance of the Linux kernel are shown in Table II. These results are essentially highly similar to those in Table I. There are two striking differences:

- The dispersion values are remarkably larger for 4 and 6 CPU cores. This is because one of the tests at 1,375,000 fps frame rate with 4 CPU cores and one of the tests at 2,000,000 fps frame rate with 6 CPU cores failed. These may be due to some relatively rare events in the system. The average is obviously affected by these outliers, which justifies our usage of median as summarizing function.
- The IPv6 throughput is slightly lower than the IPv4 throughput. There are two possible root causes for this: firstly, the frame size was larger for the IPv6 measurements, and secondly, the IPv6 addresses are four times as long as IPv4 addresses.

B. FD.io VPP Packet Forwarding Performance of R620

In order to measure the throughput of the FD.io VPP IPv4 and IPv6 packet forwarding, we used the CPU core 2 as the main core and 1 or 2 workers running on CPU cores with even or odd serial numbers to examine the performance of the test system and to gain insight into its scalability. When the CPU cores with even serial numbers were used with a single worker, we tested both core 4 and core 6 to check if there was a difference. Then they both were used with two workers. In the case of CPU cores with the odd serial numbers, the same was done with core 3, core 5, and finally, with cores 3 and 5.

At the time of our preliminary measurements using FD.io VPP with 2 workers, it was found that the Tester became the bottleneck, and thus we could not measure the true performance of the DUT. However, we considered it highly important to be able to measure the scalability of FD.io VPP at least up to two CPU cores. Therefore, the CPU clock frequency was set to 1.2 GHz (instead of 2 GHz, the nominal clock frequency of the CPU) to be able to perform the measurements using two workers.

The results of our throughput measurements of the IPv4 packet forwarding performance of FD.io VPP are shown in Table III. Core 4 and core 6 show nearly identical performance, with frame rates of 2,091,964 fps and 2,096,376 fps, respectively. When two workers were used across both cores the median frame rate roughly doubled to 4,187,904 fps. This indicates that the system scales efficiently, maximizing processing capacity without significant bottlenecks. The same can be said for the CPU cores with odd serial numbers. However, for those cores, the median value was more than 18% lower, which is clearly due to using the different NUMA node.

As for the quality of the results, with a single worker thread, the dispersion is always below 0.3%, so the results are highly stable. With two worker threads, the dispersion increased significantly, but still remained below 2%.

The results of our throughput measurements characterizing the performance of the FD.io VPP IPv6 packet forwarding are shown in Table IV. These results are basically very similar to the results in Table III. There is one visible difference: the IPv6 throughput is slightly lower than the IPv4 throughput. The possible reasons for this were explained in section A.

TABLE III
THROUGHPUT OF IPV4 PACKET FORWARDING OF THE FD.IO VPP AS A FUNCTION OF THE NUMBER AND INSTANCE OF THE ACTIVE CPU CORES, E5-2620 @ 1.2GHZ

Used CPU cores	**4th**	**6th**	**4th & 6th**	**3rd**	**5th**	**3rd & 5th**
Number of workers	**1**	**1**	**2**	**1**	**1**	**2**
median (fps)	2,091,964	2,096,376	4,187,904	1,700,940	1,697,002	3,432,319
minimum (fps)	2,089,688	2,093,309	4,183,576	1,699,461	1,695,153	3,390,624
maximum (fps)	2,093,880	2,099,060	4,238,282	1,702,287	1,698,181	3,448,609
average (fps)	2,091,944	2,096,289	4,191,721	1,700,981	1,696,889	3,431,086
standard deviation	1,501.18	1,721.36	12,034.38	849.40	726.75	13,274.60
dispersion (%)	0.20	0.27	1.31	0.17	0.18	1.69

TABLE IV
THROUGHPUT OF IPV6 PACKET FORWARDING OF THE FD.IO VPP AS A FUNCTION OF THE NUMBER AND INSTANCE OF THE ACTIVE CPU CORES, E5-2620 @ 1.2GHZ

Used CPU cores	**4th**	**6th**	**4th & 6th**	**3rd**	**5th**	**3rd & 5th**
Number of workers	**1**	**1**	**2**	**1**	**1**	**2**
median (fps)	1,919,142	1,926,082	3,876,009	1,526,365	1,587,369	3,188,186
minimum (fps)	1,916,951	1,923,811	3,866,209	1,517,536	1,585,936	3,124,999
maximum (fps)	1,921,052	1,928,529	3,908,204	1,528,194	1,588,904	3,195,313
average (fps)	1,918,898	1,926,321	3,878,199	1,525,300	1,587,528	3,185,315
standard deviation	1,120.34	1,119.19	10,495.01	3,163.36	746.77	14,657.38
dispersion (%)	0.21	0.24	1.08	0.70	0.19	2.21

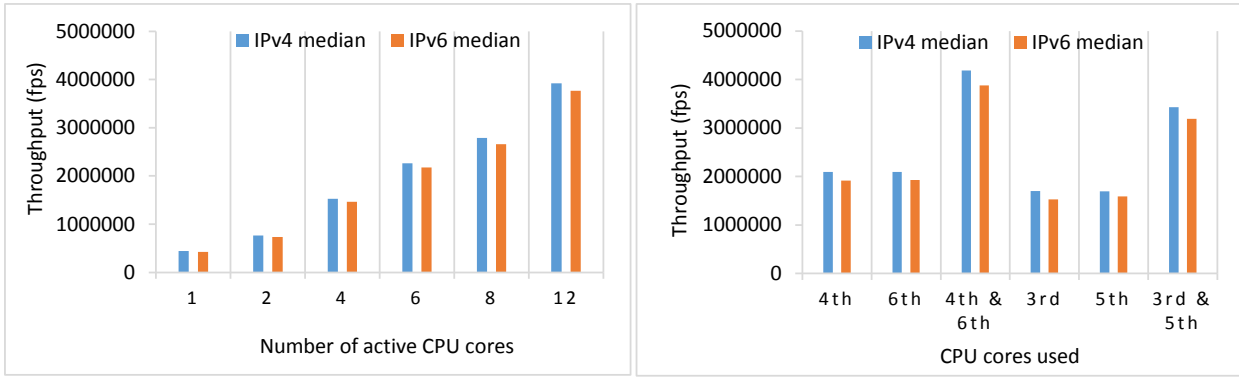


Fig. 2. Performance and scalability comparison of the Linux kernel using 1-12 active CPU cores (2GHz, left side) and FD.io VPP using one or two workers executed by the specified CPU cores (1.2GHz, right side) of a Dell PowerEdge R620 server with two 6-core Intel Xeon E5-2620 processors.

C. Comparison of the Performance of the Linux kernel and FD.io VPP using an R620 Server as DUT

The performance and scalability of the Linux kernel and FD.io VPP using a Dell PowerEdge R620 as the DUT is compared in Fig. 2. When considering their results, it is important to note that while the performance of the Linux kernel was measured at 2 GHz, the performance of FD.io VPP was measured at 1.2 GHz. Despite this disadvantage, FD.io VPP seriously outperformed the Linux kernel. While the Linux kernel on 1 CPU core delivered 442,782 fps IPv4 packet forwarding and 425,782 fps IPv6 packet forwarding performance, the FD.io VPP on 1 worker thread delivered more than 2 million IPv4 packets and more than 1.9 million IPv6 packets. Even when using CPU cores belonging to a different NUMA node than the NIC, the performance was still around 1.7 Mfps and 1.5 Mfps, for IPv4 and IPv6, respectively.

When the Linux system used all 12 CPU cores at the nominal 2 GHz clock frequency of the CPU, the performance was still only 3.92 Mfps and 3.77 Mfps for IPv4 and IPv6 packet forwarding, respectively. In contrast, FD.io VPP, using only 2 CPU cores (4 and 6), achieved more than 4.18 Mfps and 3.87 Mfps performance for IPv4 and IPv6 packet forwarding, respectively, at a clock frequency of 1.2 GHz. Our results prove that FD.io VPP is indeed a high-performance solution for IP packet forwarding.

D. Linux Kernel Packet Forwarding Performance on R730

The appropriate settings have been made using the methods described above. The same measurement parameters were used.

Each R730 server had 16 CPU cores, therefore we were able to set the active core numbers by the power of two (1, 2, 4, 8, 16). Our tests were performed using both 3.2 GHz and 1.2 GHz as the CPU clock frequency.

The IPv4 packet forwarding performance of the Linux kernel as a function of the number of active CPU cores at 3.2 GHz is shown in Table V. Using 2 CPU cores instead of 1 did not double system performance; it increased by only 1.7387 times. This phenomenon was discussed earlier. When we used 4 active CPU cores instead of 2, the performance doubled by a very good approximation (1.9293 times). The performance improvement of an 8-core system compared to 4 cores is now 1.9414 times. However, the performance increase at 16 cores was only 1.3374 times compared to that of an 8-core system. We contend that the main reason for this degradation was the insufficient performance of the NIC. To prove this, we repeated the tests by setting the CPU clock frequency to 1.2 GHz (the lowest possible value).

TABLE V
THROUGHPUT OF IPV4 PACKET FORWARDING OF THE LINUX KERNEL AS A FUNCTION OF THE NUMBER OF ACTIVE CPU CORES, E5-2667 V4 @ 3.2 GHZ

Number of CPU cores	**1**	**2**	**4**	**8**	**16**
median (fps)	746,706	1,298,271	2,504,729	4,862,591	6,503,402
minimum (fps)	744,141	1,249,991	2,499,022	4,808,590	6,492,093
maximum (fps)	750,001	1,301,758	2,509,173	4,880,386	6,516,618
average (fps)	746,678	1,295,252	2,503,804	4,859,680	6,504,297
standard deviation	1,239	10,875	3,477	21,168	7,033
dispersion (%)	0.78	3.99	0.41	1.48	0.38
relative to half as many cores	--	1.7387	1.9293	1.9414	1.3374
relative scale-up	1	0.8693	0.8386	0.8140	0.5443

TABLE VI
THROUGHPUT OF IPV4 PACKET FORWARDING OF THE LINUX KERNEL AS A FUNCTION OF THE NUMBER OF ACTIVE CPU CORES, E5-2667 V4 @ 1.2 GHZ

Number of CPU cores	**1**	**2**	**4**	**8**	**16**
median (fps)	299,758	552,915	1,093,546	2,169,302	4,046,615
minimum (fps)	298,741	548,765	1,085,934	2,160,155	4,030,760
maximum (fps)	300,659	554,762	1,095,856	2,172,245	4,056,641
average (fps)	299,694	552,454	1,093,201	2,168,249	4,045,039
standard deviation	622	1,954	1,988	3,546	6,120
dispersion (%)	0.64	1.08	0.91	0.56	0.64
relative to half as many cores	--	1.8445	1.9778	1.9837	1.8654
relative scale-up	1	0.9223	0.9120	0.9046	0.8437

TABLE VII
THROUGHPUT OF IPV6 PACKET FORWARDING OF THE LINUX KERNEL AS A FUNCTION OF THE NUMBER OF ACTIVE CPU CORES, E5-2667 V4 @ 3.2 GHZ

Number of CPU cores	1	2	4	8	16
median (fps)	728,805	1,144,471	2,233,474	4,420,087	6,505,859
minimum (fps)	718,749	1,140,602	2,124,968	4,374,999	6,374,999
maximum (fps)	730,621	1,148,438	2,239,532	4,437,501	6,562,577
average (fps)	728,192	1,143,963	2,226,907	4,417,140	6,505,697
standard deviation	2,588	2,104	2,4466	1,4143	4,0059
dispersion (%)	1,63	0,68	5,13	1,41	2,88
relative to half as many cores	--	1.5703	1.9515	1.9790	1.4719
relative scale-up	1	0.7852	0.7661	0.7581	0.5579

TABLE VIII
THROUGHPUT OF IPV6 PACKET FORWARDING OF THE LINUX KERNEL AS A FUNCTION OF THE NUMBER OF ACTIVE CPU CORES, E5-2667 V4 @ 1.2 GHZ

Number of CPU cores	1	2	4	8	16
median (fps)	288,704	511,763	1,017,832	1,990,212	3,861,507
minimum (fps)	281,249	499,999	1,007,688	1,937,483	3,749,999
maximum (fps)	289,093	512,894	1,019,907	1,996,223	3,875,001
average (fps)	288,348	511,299	1,016,904	1,988,055	3,856,157
standard deviation	1,685	2,693	3,329	12,125	25,856
dispersion (%)	2.72	2.52	1.20	2.95	3.24
relative to half as many cores	--	1.7726	1.9889	1.9553	1.9402
relative scale-up	1	0.8863	0.8814	0.8617	0.8360

The IPv4 packet forwarding performance of the Linux kernel as a function of the number of active CPU cores using 1.2 GHz CPU clock frequency is shown in Table VI. The results show that the performance of the DUT scaled up well with the increase in the number of active CPU cores in the entire range. With a 16-core system, we can observe a 1.8654 times increase compared to 8 cores. On one hand, this is certainly an improvement (especially compared to 1.3374); however, it is still lower than the increases seen when scaling from 2 to 4 or from 4 to 8 cores in the same test series, despite the NIC capacity being higher than the measured throughput. We attribute this small degradation to the fact that the 16 cores were competing to access the NIC. Overall, the system performance scaled up well: when using 16 CPU cores instead of 1 CPU core, the throughput increased from 299,758 fps to 4,046,615 fps by a factor of 13.5.

The IPv6 packet forwarding performance of the Linux kernel as a function of the number of active CPU cores using 3.2 GHz CPU clock frequency is shown in Table VII. These results are basically very similar to those in Table V with the difference that IPv6 throughput is slightly lower than IPv4 throughput.

Although IPv6 throughput is generally slightly lower than IPv4, the median values for 16 CPU cores are very similar (6,503,402 fps for IPv6 and 6,505,859 fps for IPv4). This suggests a hardware limitation, likely the network interface, in this case.

The IPv6 packet forwarding performance of the Linux kernel as a function of the number of active CPU cores using 1.2 GHz CPU clock frequency is shown in Table VIII. Overall, the system performance scaled well: when using 16 CPU cores instead of 1 CPU core, the throughput increased from 288,704 fps to 3,861,507 fps by a factor of 13.4.

E. FD.io VPP Packet Forwarding Performance on R730

The appropriate settings have been made using the methods described above. The measurement parameters used so far were also used for these measurements.

In order to measure the throughput of the FD.io VPP IPv4 and IPv6 packet forwarding, we used the CPU core 2 as the main core and 1 or 2 workers running on CPU cores with even serial numbers to examine the performance of the test system and to gain insight into its scalability. Previously, we also used CPU cores with odd serial numbers to examine the test system's performance. However, as discussed, there were no relevant differences in the results, aside from being lower due to their association with a different NUMA node than the NIC. This time, we focused on providing a clear comparison between different clock speeds (3.2 GHz and 1.2 GHz) and omitted the use of CPU cores with odd serial numbers.

The results of our throughput measurements of the IPv4 packet forwarding performance of FD.io VPP are shown in Table IX. Examining core 4 and core 6 at 3.2 GHz, we see that there is no significant difference in performance (6,947,546 fps and 6,960,875 fps) but compared to the results measured at 1.2 GHz (2,886,634 fps and 2,887,674 fps), there is a nearly 2.5-fold increase. For two workers, we can say that our results approximately doubled when comparing the single worker and the two worker results at 1.2 GHz.

As for the quality of the results, they are highly stable and consistent because all of the dispersions are below 1%.

The results of the throughput measurements of the IPv6 packet forwarding performance of FD.io VPP are shown in Table X. These results are basically very similar to the results in Table IX with the difference that IPv6 throughput is slightly lower than IPv4 throughput. Overall, we have highly stable results because all of the dispersions are below 1%.

It is salient that the results with FD.io VPP are more stable than those with the Linux kernel. The reason behind this is the following: during FD.io VPP measurements, the CPUs used for executing the workers were isolated (using the **`isolcpus`** kernel command line parameter). This means that no other task could be scheduled to the isolated CPU cores by the kernel. Conversely, the Linux kernel used all CPU cores for packet forwarding and the scheduler occasionally assigned them other tasks, as well.

TABLE IX
THROUGHPUT OF IPV4 PACKET FORWARDING OF THE FD.IO VPP AS A FUNCTION OF THE NUMBER AND INSTANCE OF THE ACTIVE CPU CORES, E5-2667 V4

Used CPU cores	**4th**	**6th**	**4th**	**6th**	**4th & 6th**
Number of workers	**1**	**1**	**1**	**1**	**2**
CPU clock frequency	**3.2 GHz**	**3.2GHz**	**1.2 GHz**	**1.2 GHz**	**1.2 GHz**
median (fps)	6,947,546	6,960,875	2,886,634	2,887,674	5,866,613
minimum (fps)	6,942,869	6,937,499	2,874,999	2,885,736	5,859,374
maximum (fps)	6,953,156	6,968,751	2,887,939	2,888,488	5,869,027
average (fps)	6,948,498	6,959,671	2,886,116	2,887,500	5,866,099
standard deviation	3,077	6,674	2,906	747	2,680
dispersion (%)	0.15	0.45	0.45	0.10	0.16

TABLE X
THROUGHPUT OF IPV6 PACKET FORWARDING OF THE FD.IO VPP AS A FUNCTION OF THE NUMBER AND INSTANCE OF THE ACTIVE CPU CORES, E5-2667 V4

Used CPU cores	**4th**	**6th**	**4th**	**6th**	**4th & 6th**
Number of workers	**1**	**1**	**1**	**1**	**2**
CPU clock frequency	**3.2 GHz**	**3.2GHz**	**1.2 GHz**	**1.2 GHz**	**1.2 GHz**
median (fps)	6,250,366	6,253,601	2,559,248	2,555,627	5,215,632
minimum (fps)	6,218,749	6,240,721	2,546,874	2,554,662	5,187,499
maximum (fps)	6,258,057	6,264,577	2,559,815	2,556,732	5,220,337
average (fps)	6,251,035	6,252,592	2,558,535	2,555,526	5,213,860
standard deviation	3,195	4,308	3,023	666	3,670
dispersion (%)	0.63	0.38	0.51	0.08	0.63

F. Comparison of the performance of the Linux kernel and FD.io VPP using an R730 Server as DUT

Unlike in the case of the R620 test system, the results of the Linux kernel and the FD.io VPP using the same CPU clock frequency on the R730 test system are directly comparable.

First, we compare the results produced using 1.2 GHz CPU clock frequency. The results are shown in Fig. 3. Whereas the Linux kernel on 1 CPU core delivered 299,758 fps IPv4 packet forwarding and 288,704 fps IPv6 packet forwarding performance, the FD.io VPP with 1 worker thread delivered more than 2.8 million IPv4 packets and more than 2.5 million IPv6 packets. When the Linux system used all 16 CPU cores of the CPU, the performance was only 4.05 Mfps and 3.86 Mfps for IPv4 and IPv6 packet forwarding, respectively. In contrast, FD.io VPP, using only 2 CPU cores (4 and 6), achieved 5.87 Mfps and 5.22 Mfps performance for IPv4 and IPv6 packet forwarding, respectively.

The 3.2 GHz results are compared in Fig. 4. The single core system of Linux kernel delivered 746,706 fps IPv4 packet forwarding and 728,805 fps IPv6 packet forwarding performance, whereas the FD.io VPP on 1 worker thread performance, whereas the FD.io VPP on 1 worker thread delivered more than 6.95 million IPv4 packets and more than 6.25 million IPv6 packets. Our results prove that FD.io VPP is indeed a high-performance solution for IP packet forwarding.

G. Comparison of the results of the R620 and R730 Servers

When comparing the two types of servers, it is important to note that the nominal CPU clock frequencies of the R620 and R730 servers are 2 GHz and 3.2 GHz, respectively.

When comparing the median values of their Linux kernel IPv4 packet forwarding results measured at their nominal CPU frequencies, a single CPU core of the R730 server outperformed the single CPU core of the R620 server with a factor of 1.6864 (746,706 fps vs. 442,782 fps) and similar statements can be made regarding their performances from 2 to 8 CPU cores. Alternatively, it can be noted that the performance of a single core of the E5-2667 v4 CPU at 3.2 GHz (746,706 fps) is nearly equivalent to the performance of two cores of the E5-2620 CPU at 2 GHz (765,223 fps).

As for the FD.io VPP results of the two types of CPUs, those measured at 1.2 GHz are directly comparable. The 2.89 Mfps

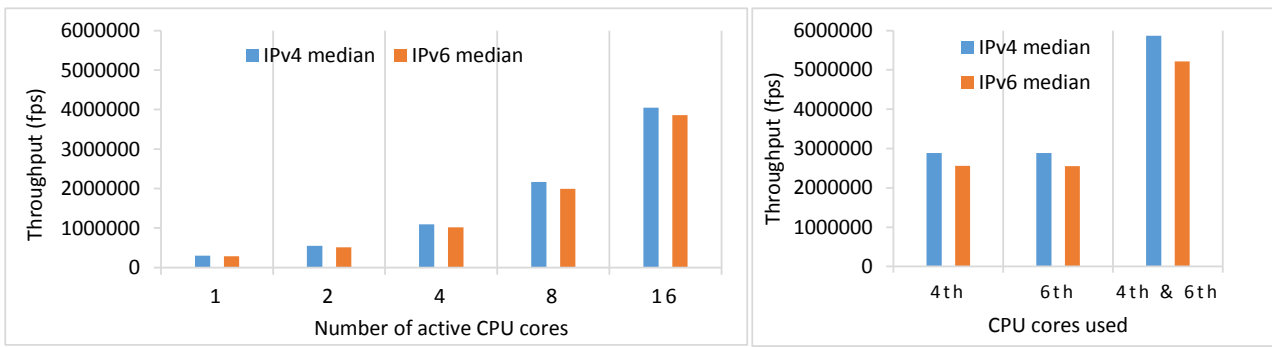


Fig. 3. Performance and scalability comparison of the Linux kernel using 1-16 active CPU cores (left side) and FD.io VPP using one or two workers executed by the specified CPU cores (right side) of a Dell PowerEdge R730 server with two 8-core Intel Xeon E5-2667 v4 processors @ 1.2GHz.

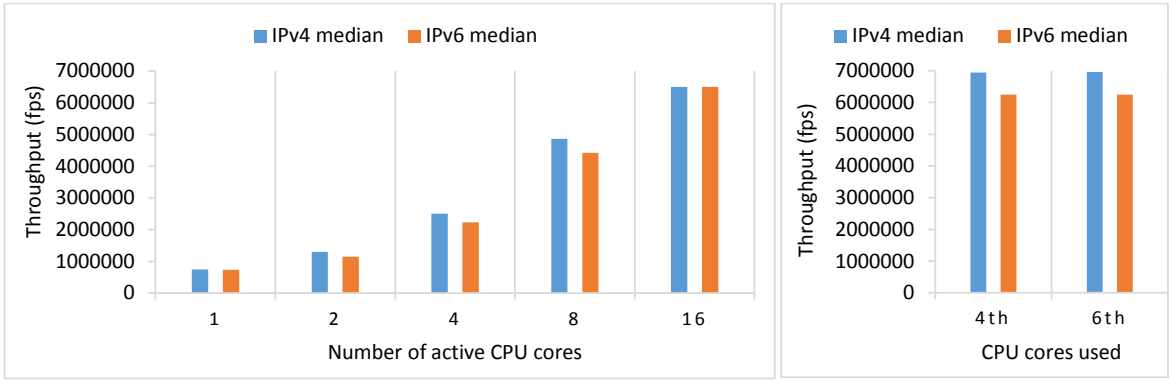


Fig. 4. Performance and scalability comparison of the Linux kernel using 1-16 active CPU cores (left side) and FD.io VPP using one worker executed by the specified CPU cores (right side) of a Dell PowerEdge R730 server with two 8-core Intel Xeon E5-2667 v4 processors @ 3.2GHz.

single core IPv4 throughput of the E5-2667 v4 CPU @ 1.2 GHz shows a 1.38-fold increase compared to the 2.09 Mfps single core IPv4 throughput of the E5-2620 CPU @ 1.2 GHz. The higher nominal clock frequency of the newer CPU further amplifies this difference.

VII. Discussion and Future Research

In our research group's previous benchmarking effort, the number of CPU cores in the test system was equal to powers of 2 (see [19]). This approach allowed us to consistently double the number of active CPU cores in each iteration and fully utilize all available CPU cores in the final test. During our first attempt to compare the performance of the Linux kernel and FD.io VPP, only servers with 12 CPU cores were available for the task. For this reason, we also conducted tests using 6 cores to provide a basis for comparison with the tests with 12 cores. Building on the success of our initial effort [6], we decided to repeat our experiments with 16-core servers, too.

Our current measurements only included throughput tests. In the future, we also plan to examine the latency. Since latency is measured at the frame rate previously determined by the throughput tests, our hypothesis is that FD.io VPP will exhibit higher latency. This is due to its vector packet processing mechanism, where a substantial number of frames are first accumulated into a vector before being processed together as they traverse the nodes of the packet processing graph. The latency results will complement the throughput results to give a more comprehensive view of the performance of the tested IP packet forwarding solutions.

Since both FD.io VPP and siitperf use DPDK, and siitperf only uses one core for sending and one core for receiving in each direction, we could evaluate the performance of FD.io VPP only up to two working threads. This remains the case even when using the trick of lowering the CPU clock frequency of the DUT. To be able to work with a significantly higher number of workers (4, 8, etc.) a more powerful Tester would be needed. The long-term plan of our research group includes the building of an FPGA-based tester that implements the functionalities of siitperf.

Another interesting research direction could be to evaluate the performance of the Linux kernel when using NAPI polling mode [20].

It would also be worth comparing the performance of FD.io VPP to that of Open vSwitch with DPDK (OvS DPDK) and eXpress Data Path (XDP).

Our results will encourage network operators to use FD.io VPP in production networks for IPv4 and IPv6 packet forwarding. Exhibiting high performance when running on commodity servers and being free software, FD.io VPP can be a good alternative to commercial routers. However, high performance and low cost are only two aspects. Network operators must consider at least two other ones: security and support. FD.io VPP runs on Linux, a general-purpose and widely used operating system. It is crucial to ensure the secure operation of the host machine. To that end, it must be carefully configured, and security updates must be regularly installed. Major version upgrades to the operating system could involve issues when DPDK and FD.io versions are upgraded. Fortunately, Debian and Ubuntu have long-term support (LTS) versions; thus, they can be operated for several years without changing the major version. Router vendors also provide support. As free software is basically provided “as is” without support, network operators either need to employ experts or buy support from a company that employs experts in Linux, DPDK and FD.io VPP.

VIII. Conclusion

We measured the performance of IPv4 and IPv6 packet forwarding of the Linux kernel and FD.io VPP. Regardless of the IP version, the Linux kernel’s packet forwarding performance showed a good scale-up with the increasing number of CPU cores, although minor performance variations were observed when utilizing different NUMA nodes. In contrast, FD.io VPP tests demonstrated exceptionally high performance with perfect scalability from 1 to 2 workers.

When a Dell PowerEdge R620 server with two 6-core E5-2620 CPUs was used as the DUT, FD.io VPP, using 2 workers executed by two CPU cores running at 1.2 GHz, outperformed

the Linux kernel, using 12 CPU cores running at 2 GHz.

When a Dell PowerEdge R730 server with two 8-core E5-2667 v4 CPUs was used as the DUT, the same two CPU clock speeds were used for testing both solutions. At 1.2 GHz CPU clock frequency, the Linux kernel on 1 CPU core delivered 299,758 fps IPv4 packet forwarding and 288,704 fps IPv6 packet forwarding performance, and the FD.io VPP with 1 worker thread delivered more than 2.8 million IPv4 packets and more than 2.5 million IPv6 packets. When the Linux system used all 16 CPU cores of the CPU, the performance was only 4.05 Mfps and 3.86 Mfps for IPv4 and IPv6 packet forwarding, respectively. In contrast, FD.io VPP, using only 2 CPU cores (4 and 6), achieved 5.87 Mfps and 5.22 Mfps performance for IPv4 and IPv6 packet forwarding, respectively. At a CPU clock frequency of 3.2 GHz, the Linux kernel's single-core system achieved an IPv4 packet forwarding performance of 746,706 fps and an IPv6 performance of 728,805 fps. In comparison, FD.io VPP, using a single worker thread, delivered over 6.95 million IPv4 packets and more than 6.25 million IPv6 packets. Our results confirm that FD.io VPP is indeed a high-performance solution for IP packet forwarding.

ACKNOWLEDGEMENT

The authors thank Bertalan Kovács for reviewing and commenting on the manuscript.

The authors thank Natasha Bailey-Borbély, Széchenyi István University, for the English language proofreading of the manuscript.

REFERENCES

[1] FD.io "What is the Vector Packet Processor (VPP)", VPP official website, https://s3-docs.fd.io/vpp/24.06/

[2] S. Bradner, and J. McQuaid, "Benchmarking methodology for network interconnect devices", IETF RFC 2544, 1999. DOI: 10.17487/RFC2544.

[3] D. Newman, T. Player, "Hash and stuffing: Overlooked factors in network device benchmarking", IETF RFC 4814, 2008. DOI: 10.17487/RFC4814.

[4] T. Herbert, W. de Bruijn, "Scaling in the Linux Networking Stack", https://www.kernel.org/doc/Documentation/networking/scaling.txt

[5] C. Popoviciu, A. Hamza, G. V. de Velde, and D. Dugatkin, "IPv6 benchmarking methodology for network interconnect devices", IETF RFC 5180, 2008, DOI: 10.17487/RFC5180.

[6] M. Kosák, and G. Lencse, "Performance comparison of IP packet forwarding solutions", *2024 47th International Conference on Telecommunications and Signal Processing (TSP)*, Virtual Conference, July 10-12, 2024, pp. 243–248. DOI: 10.1109/TSP63128.2024.10605773

[7] G. Slavic, N. Krajnovic, "Practical implementation of the vector packet processing software router", *2024 32nd Telecommunications Forum (TELFOR)*, Crowne Plaza, Belgrade, Serbia, November 26-27, 2024, DOI: 10.1109/TELFOR63250.2024.10819057

[8] Y. Peng, Y. Xiao, J. Duan, X. Zhang and W. Li, "Emulating high-performance networks with CNNet", *2023 IEEE 43rd International Conference on Distributed Computing Systems Workshops (ICDCSW)*, Hong Kong, Hong Kong, 2023, pp. 91–96, DOI: 10.1109/ICDCSW60045.2023.00024.

[9] S. Han, S. Wang, Y. Huang, T. Huang, Y. Liu, "High-performance and low-cost VPP gateway for virtual cloud networks" *2022 IEEE Global Communications Conference (GLOBECOM)*, pp. 4419–4424, DOI: 10.1109/GLOBECOM48099.2022.10001162

[10] P. Cai, M. Karsten, "Kernel vs. user-level networking: Don't throw out the stack with the interrupts", *ACM SIGMETRICS Performance Evaluation Review*, vol. 52, no. 1, June 13, 2024, pp. 43–44, DOI: 10.1145/3673660.3655061

[11] DPDK official website, https://www.dpdk.org/about/

[12] B. Jacob, S. W. Ng, and D. T. Wang. *Memory Systems: Cache, DRAM, Disk*, Morgan Kaufmann Publishers, 2008. DOI: 10.1016/B978-0-12-379751-3.X5001-2.

[13] G. Lencse, "Design and implementation of a software tester for benchmarking stateless NAT64 gateways", *IEICE Transactions on Communications*, vol. E104-B, no. 2, pp. 128–140, February 2021, DOI: 10.1587/transcom.2019EBN0010.

[14] M. Georgescu, L. Pislaru, and G. Lencse, "Benchmarking methodology for IPv6 transition technologies", IETF RFC 8219, August 2017, DOI: 10.17487/RFC8219.

[15] G. Lencse, "Adding RFC 4814 random port feature to siitperf: Design, implementation and performance estimation", *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, vol. 9, no. 3, 2020, DOI: 10.11601/ijates.v9i3.291.

[16] G. Lencse, "Checking the accuracy of siitperf", *Infocommunications Journal*, vol. 13, no. 2, pp. 2–9, June 2021, DOI: 10.36244/ICJ.2021.2.1

[17] G. Lencse, "Design and implementation of a software tester for benchmarking stateful NATxy gateways: Theory and practice of extending siitperf for stateful tests", *Computer Communications*, vol. 172, no. 1, pp. 75–88, Aug. 1, 2022, DOI: 10.1016/j.comcom.2022.05.028.

[18] G. Lencse, "Making stateless and stateful network performance measurements unbiased", *Computer Communications*, vol. 225, September 2024, pp. 141–155, DOI: 10.1016/j.comcom.2024.05.018

[19] G. Lencse, and Á. Bazsó, "Benchmarking methodology for IPv4aaS technologies: Comparison of the scalability of the Jool implementation of 464XLAT and MAP-T", *Computer Communications*, vol. 219, April 2024, pp. 243–258, DOI: 10.1016/j.comcom.2024.03.007

[20] J. Sahoo, "Deep dive into NAPI: Optimizing network performance in the Linux kernel", March 3, 2024, https://www.linkedin.com/pulse/deep-dive-napi-optimizing-network-performance-linux-kernel-sahoo-opa1c/

Melinda Kosák received her BSc in electrical engineering from Széchenyi István University, Győr, Hungary in 2024.

She has been a member of the Cybersecurity and Network Technologies Research Group of the Faculty of Mechanical Engineering, Informatics and Electrical Engineering of Széchenyi István University, Győr, Hungary since 2022. Her research interests include performance analysis of FD.io VPP. She is now continuing her studies at Széchenyi István University as an MSc student in electrical engineering.

Gábor Lencse received his MSc and PhD degrees in computer science from the Budapest University of Technology and Economics, Budapest, Hungary in 1994 and 2001, respectively.

He has been working for the Department of Telecommunications, Széchenyi István University, Győr, Hungary since 1997 and has attained the rank of Professor. He has also been a part-time Senior Research Fellow at the Department of Networked Systems and Services, Budapest University of Technology and Economics since 2005. His research interests include the performance and security analysis of IPv6 transition technologies. He is a co-author of RFC 8219 and RFC 9313.

Multipath Rate Control for Real-Time Media

Balázs Kreith and Árpád Drozdy

Abstract—**In this paper, we present rate control algorithms designed for real-time media transmission over multiple paths. Our focus is on delivering media content simultaneously across multiple paths while maximizing the network traffic utilization on each path. The proposed algorithms ensure reduced network delay fluctuations, more consistent transmission rates for media content, and fairness to cross-traffic across all paths. To evaluate their performance, the solution is implemented and tested in an emulated networking environment under various test scenarios. The results demonstrate that applying our algorithms leads to reduced network delay fluctuations, improved structural similarity of the received media content, and enhanced fairness toward cross-traffic.**

Index Terms—**Multipath, RTP, Multimedia, Scheduling, Congestion Control, Real-Time Media**

I. INTRODUCTION

Web Real-time Communication (WebRTC) has enabled the widespread deployment of real-time multimedia communications on the Internet. Today, videoconferencing platforms (e.g., Google Meet, MS Teams, Discord) and cloud gaming services (e.g., GeForce NOW) have become integral to daily life. Most of these real-time communication (RTC) services rely on WebRTC [1].

The rapid growth of network traffic has driven efforts to improve transport protocols for multimedia systems, focusing on maximizing bandwidth utilization while avoiding packet losses caused by exceeding end-to-end path capacity.

Conversational multimedia communications impose strict latency limits to maintain the recommended mouth-to-ear delay of 150ms, ensuring fluent real-time conversations [2]. Unlike bulk data transfer, loss detection and retransmission may not be feasible in these scenarios, as delayed packets risk missing their playout deadlines at the receiver.

To address the challenges of reliable congestion control for real-time multimedia, the Internet Engineering Task Force (IETF) previously addressed this issue through the Real-Time Protocol Media Congestion Avoidance Technique (RMCAT) working group [3]. RMCAT introduced several congestion control algorithms [4]–[6], primarily designed for single-path delivery, with minimal focus on multipath scenarios.

Modern devices frequently feature multiple network interfaces, enabling endpoints to communicate via multiple, potentially disjoint paths. These paths can be leveraged for load balancing, capacity aggregation, or failover in case of path failures.

Balázs Kreith is with the Department of Telecommunications, University of Debrecen, Debrecen, Hungary. (E-mail: balazs@kreith.hu).
Árpád Drozdy is with Nokia, Budapest, Hungary. (E-mail: arpad.drozdy@nokia.com).

DOI: 10.36244/ICJ.2025.2.6

The Real-time Transport Protocol (RTP) [7], widely used for real-time media, is supported by the Real-time Transport Control Protocol (RTCP) for exchanging performance metrics. To extend RTP to multipath scenarios, Multipath RTP (MPRTP) [8] was introduced. While MPRTP provides mechanisms for shifting traffic between congested and non-congested paths, it does not fully explore or maximize the end-to-end path capacities. Achieving this requires individual congestion control algorithms for each path and a mechanism to coordinate them effectively.

In this paper, we propose a multipath rate control algorithm designed to fully exploit multiple paths for real-time media. Our solution achieves lower network delay fluctuations, improved media content similarity, and fairness for cross-traffic.

We introduce an architecture that builds on existing RMCAT congestion control algorithms enabling efficient multipath utilization. The novelty of our approach lies in coupling congestion control algorithms across paths to ensure fairness, prevent traffic fading during rebalancing, and optimize real-time media performance.

The remainder of this paper is organized as follows: Section II outlines our design goals for a multipath rate controller. Section III presents a high-level architecture compatible with existing congestion control algorithms. Section IV details our proposed algorithms, which are evaluated under various scenarios in Section V. Section VI discusses implementation considerations, Section VII reviews related work on congestion control in multipath communication, and Section VIII concludes the paper.

II. DESIGN GOALS

Our goal is to deliver real-time media over multiple paths simultaneously. To achieve this, we introduce a rate control mechanism that *1)* distributes media content across multiple paths simultaneously, and *2)* maximizes the network traffic sent over these paths.

A. Delivering media content over multiple paths simultaneously

In multi-path media transmission, packets are distributed across *subflows* (paths) as shown in Figure 1. These paths may have varying delays and jitter, requiring synchronization at the receiver. To manage this, we designed a **Multipath Packet Processing Module** (MPPM), which:

- Distributes packets based on allocated bitrates.
- Minimizes outbound traffic fluctuations to avoid self-inflicted congestion.
- Prevents traffic fading, where traffic exceeds a path's capacity due to unsynchronized bitrate changes.

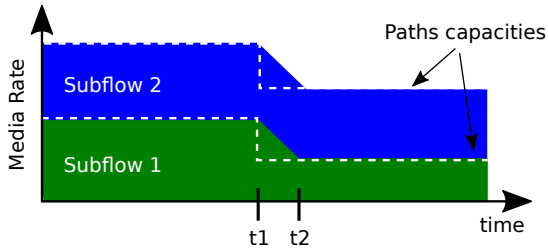


Fig. 1. Two subflows are shown. At $t1$, congestion is detected on $Subflow\ 1$, and the allocated bitrate is reduced. The media encoder adjusts to the new allocation, reaching the demanded bitrate at $t2$. Without synchronization, $Subflow\ 2$ would receive more traffic than its capacity between $t1$ and $t2$.

The MPPM updates packet distribution ratios promptly but must synchronize with the media encoder, which requires time to adjust. This synchronization avoids congestion or underutilization caused by temporary bitrate mismatches.

B. Maximizing the network traffic sent on paths

To maximize network utilization, a congestion control algorithm estimates and adjusts bitrates for each path. In single-path communication, the estimated bitrate equals the actual bitrate. For multi-path communication, separate estimations are needed for each path.

In our solution, the **Multipath Rate Control Module** (MRCM) ensures fairness across paths, following the fairness principles of MPTCP [9]. Fairness balances bandwidth across all paths and flows. To achieve this, the MRCM *1)* dynamically adjusts bitrates when congestion is detected, and *2)* increases bitrates when higher capacity is estimated, while maintaining fairness for cross-traffic.

Simulations and tests demonstrate how the MPPM avoids traffic fading and how the MRCM achieves fairness and efficiency in multi-path scenarios (see Sections V-A and V-B).

III. Architecture

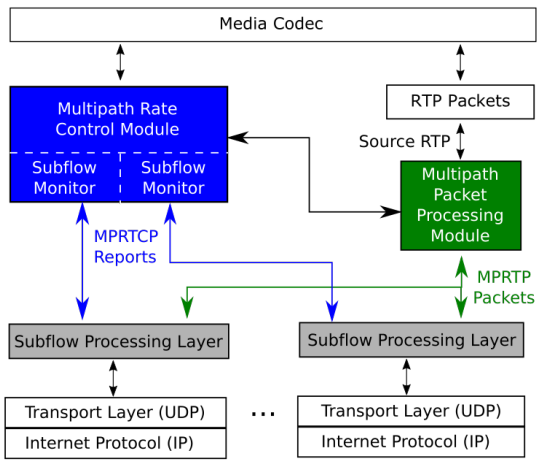


Fig. 2. Architectural scheme for Media Systems using Multipath Rate Control.

We propose a layered architectural model for a complex multimedia system (Figure 2), incorporating two modules to implement the multipath rate control mechanism: the Multipath Rate Control Module (MRCM) and the Multipath Packet Processing Module (MPPM). These modules operate between the layers responsible for media content and network transmissions. Additionally, Subflow Monitor (SM) submodules are applied to all subflows under MRCM's control. Note that the module implementations differ between the sender and receiver sides. For sending and receiving real-time media packets over multiple paths, we use Multipath Real-Time Protocol (MPRTP) [8], with subflow reports exchanged via Multipath Real-Time Control Protocol (MPRTCP).

To clarify the workings of the modules, we define the key types of sending bitrates used throughout the paper. The sending bitrate, representing the number of bits transmitted per second on a path, directly impacts media quality. In multi-path connections, the sender dynamically adjusts the sending bitrate for each path based on capacity estimations. The modules operate with the bitrate definitions presented in Table I as shown in Figure 3.

TABLE I
Definitions of Key Bitrate Metrics.

Metric	Notation	Description
Estimated bitrate	$SR_{estimated}$	Predicted bandwidth capacity of a link.
Actual bitrate	SR_{actual}	The bitrate currently transmitted on a path.
Allocated bitrate	$SR_{allocated}$	The bitrate calculated and assigned to a path.
Stable bitrate	SR_{stable}	The achieved and maintained bitrate by the media encoder.
Transient bitrate	$SR_{transient}$	A temporary bitrate during the encoders transition to a new allocation.

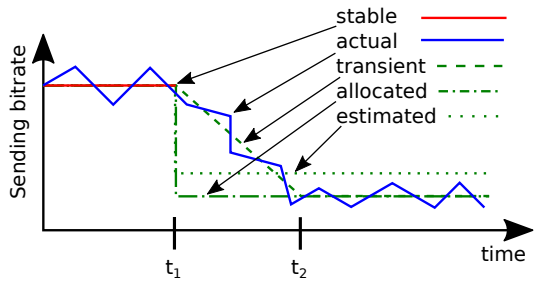


Fig. 3. Different types of sending bitrates during a reduction phase.

A. Subflow Monitor Module

Subflows can have varying characteristics, such as jitter, RTT, and bandwidth capacities, which may change during the connection. Therefore, each subflow is monitored by a *congestion control algorithm*, which evaluates the path characteristics and calculates the $SR_{estimated}$ for the monitored subflow. This information is forwarded to the MRCM on the sender side.

In addition to the $SR_{estimated}$, the SM submodule also informs the MRCM of the path state for each subflow. The path state can be *Congested*, *Stable*, or *Underused*. The MRCM uses this information to track the SR_{stable} when the stable state is reported. This state information is also utilized by the MPPM to avoid congested paths when possible. Based on the path state, the MPPM may assign packets to subflows, ensuring that, for example, I-Frame (reference points for decoding subsequent frames) packets are not sent on paths marked as *Congested*.

B. Multipath Rate Control Module

The MRCM operates only on the sender side. It sends $SR_{allocated}$, SR_{stable}, and path states to the MPPM. The MRCM calculates $SR_{allocated}$ from the subflows' $SR_{estimated}$ and SR_{stable} based on the following principles:

1) If $SR_{estimated} \leq SR_{stable}$ for a subflow, the algorithm assumes congestion and sets $SR_{allocated}$ equal to $SR_{estimated}$.
2) If $SR_{estimated} > SR_{stable}$ the algorithm assumes ramp-up and regulates $SR_{allocated}$ to ensure fairness, depending on how many subflows ramp-up simultaneously.

When a subflow reduces its $SR_{estimated}$, $SR_{allocated}$ is set to the new estimation to end congestion. Conversely, during ramp-up, $SR_{allocated}$ is less than or equal to $SR_{estimated}$, depending on the number of subflows ramping up. If only one subflow ramps-up, $SR_{allocated}$ equals $SR_{estimated}$. If multiple subflows ramp-up, $SR_{allocated}$ is reduced. If the total bitrate growth of all subflows does not exceed the maximum achievable growth of a single-path flow, fairness is ensured [9].

A key question is which subflows should determine the bitrate growth. The MRCM selects one subflow at a time and distributes its bitrate growth among all subflows. The MRCM executes the *fractional distribution algorithm* whenever a new $SR_{estimated}$ is reported, calculating the updated $SR_{allocated}$, which is sent to the MPPM with the SR_{stable} and path state.

C. Multipath Packet Processing Module

At the sender side, the MPPM distributes incoming traffic based on the information from the MRCM. This helps avoid traffic fading, as discussed in section II. Additionally, the MPPM considers real-time boundaries by selecting different paths for I- and B-Frames when possible. To assign a subflow to a packet, the MPPM runs the *packet scheduler algorithm* upon receiving a packet.

The *packet scheduler algorithm* converts RTP packets to MPRTP packets and assigns them to available subflows. It selects only subflows with a path state of *Stable* or *Underused*. Note that packet type information can be extracted only if the packet is not encrypted at the time of receipt, as discussed in section VI.

When subflows' $SR_{allocated}$ change, the media encoder responds with a time lag. This lag must be considered to avoid fluctuations that could lead to traffic fading. The scheduler calculates a $SR_{transient}$ for each subflow at every point in time when a packet is mapped, ensuring the distribution ratio is based on $SR_{transient}$ to allow seamless bitrate transitions. The packet scheduler algorithm is described in the next section.

Note that packets should only be sent through a subset of subflows if their SR_{actual} is smaller than $SR_{transient}$. Otherwise, fluctuations around $SR_{transient}$ are too high, increasing the risk of traffic fading.

At the receiver side, incoming MPRTP packets from different subflows are combined and played out to the Media Codec. A dejitter buffer designed for MPRTP packets from multiple subflows, as presented in [8], is used in our implementation.

IV. ALGORITHMS

To deliver media over multiple paths, algorithms are developed for the modules in section III. The MRCM uses the fractional distribution algorithm to allocate sending-bitrates for subflows, with each subflow monitored by a Subflow Monitor (SM) applying a congestion control algorithm. The MPPM uses the packet scheduler algorithm to map incoming packets to subflows based on allocated bitrates.

A. Congestion Control Algorithm

Each subflow is monitored by its own SM, which applies the FRACTaL [10] rate control algorithm. FRACTaL calculates the sending-bitrate for the path, which is forwarded as the $SR_{estimated}$ for the subflow. FRACTaL uses a state machine with REDUCE, KEEP, PROBE, and INCREASE states, mapped by the SM to path states defined in section III. If the inner state is REDUCE or INCREASE, the reported path state is *Congested* or *Underused*, respectively. Otherwise, the state is *Stable*, and the SR_{stable} is updated.

The SM sends the $SR_{estimated}$, SR_{stable}, and path state to the MRCM, which executes the fractional distribution algorithm upon receiving new information from any SM.

B. Fractional Distribution Algorithm

Algorithm 1 Fractional Distribution for Each Subflow

Require: The $total\_ramp\_up\_rate$ bitrate must be distributed among the subflows.
Require: An array of $weights$ contains a weight for each ramping-up subflow.
Require: The $total\_weight$, which is the sum of weights of all ramping-up subflows.

```
M̂R ← 0
subflows_total ← len(subflows)
for j = 1 → subflows_total do
    if ER[j] ¡ AR[j] then
        AR[j] ← ER[j]
    else
        ramp_up = total_ramp_up × weights[j]/total_weight
        if SR[j] ¡ AR[j] then
            ramp_up = MIN(AR[j] − SR[j], ramp_up)
        AR[j] ← SR[j] + ramp_up
    M̂R ← M̂R + AR[j]
```

The fractional distribution algorithm calculates $SR_{allocated}$ for each subflow, based on the principles

from section III-B. It takes $SR_{estimated}$, SR_{stable}, and path state from each subflow, reported by the Subflow Monitors.

If $SR_{estimated}$ is less than or equal to SR_{stable}, $SR_{allocated}$ is set to $SR_{estimated}$. Otherwise, let ΔSR be the positive difference between $SR_{estimated}$ and SR_{stable}. The algorithm selects the subflow with the highest ΔSR, and sets the $total\_ramp\_up\_rate$ to the highest ΔSR. The total rate increase of the ramping-up subflows should not exceed this rate to ensure fairness for cross-traffic, as described in section II.

A $weights$ array is computed, where each element is the ratio of the subflows ΔSR and Round Trip Time (RTT), allowing subflows with lower RTTs to ramp-up faster. $total\_weight$ is the sum of the $weights$. In practice, an encoder may not provide a new $SR_{allocated}$ immediately, leading to potential overlap of multiple ramp-ups. If a new $SR_{allocated}$ for a ramping-up subflow is smaller than the previous one, the smaller value is used.

The fractional distribution algorithm is shown in Algorithm 1. Note that fairness across multiple paths depends on each congestion control algorithm being fair on a single path, a topic beyond the scope of this paper. FRACTaL is a proactive congestion control algorithm that uses adaptive thresholds to detect congestion and compete with loss-based algorithms.

C. Packet Scheduler Algorithm

Algorithm 2 Map a Subflow to a Packet

Require: The $target^{(off)}$ value
Require: The $selected\_subflows$ array
  $selected \leftarrow NULL$
  $selected\_diff \leftarrow 0$
  for $j = 1 \rightarrow selected\_subflows_{total}$ **do**
    $target \leftarrow AR[j] \times target^{(off)} + SR[j] \times (1 - target^{(off)})$
    $diff \leftarrow target - SentBitrate[j]$
    if $selected$ **is** $NULL$ **OR** $selected\_diff$ ¡ $diff$ **then**
      $selected \leftarrow subflow$
      $selected\_diff \leftarrow diff$
  return $selected$

The MPPM executes the packet scheduler algorithm each time it receives a packet. It selects a subflow based on the packet size, $SR_{allocated}$, SR_{stable}, and the path state of each subflow. To avoid traffic fading, a $SR_{transient}$ is calculated as a weighted average of $SR_{allocated}$ and SR_{stable}, where the weight, $target^{(off)}$, is determined by how close MR_{actual} is to $MR_{demanded}$:

$$target^{(off)} = 1 - \left| \frac{MR_{demanded} - MR_{actual}}{MR_{demanded} - \sum SR_{stable}} \right| \quad (1)$$

When $target^{(off)}$ is 0, MR_{actual} has reached $SR_{allocated}$. The packet scheduler maps the subflow to a packet with the largest difference between its $SR_{transient}$ and SR_{actual}. The packet scheduler algorithm is presented in Algorithm 2.

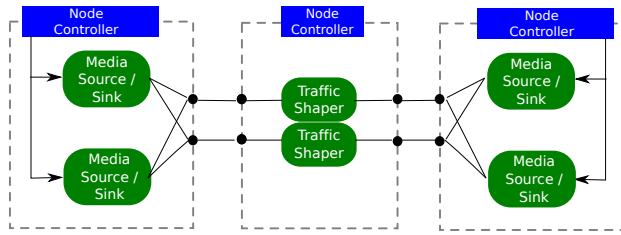


Fig. 4. Performance Evaluation Setup.

V. PERFORMANCE EVALUATION

In this section, we evaluate the performance of our proposal using multiple paths for real-time multimedia communications. Our algorithms are implemented with Gstreamer [11] and can be used with OpenWebRTC [12]. We test the algorithms effectiveness in avoiding traffic fading and ensuring fairness to cross-traffic.

Environment: The tests are conducted in a controlled environment (Figure 4), where the sender and receiver run on separate Linux machines connected via Ethernet. Network conditions are emulated with Netem [13] and tbf [14]. Traffic shapers set a 300 ms queue length with a drop-tail policy, and network delays in bottleneck links are 100 ms. We use the KristenAndSara [15] video sequence with VP8 [16], simulating a camera source at a clock rate of 90000 and a frame rate of 25 fps. Full details are available online[1].

Metrics: We measure Goodput (GP), Packet Loss Rate (LR), and Average Bandwidth Utilization (ABU). Queue Median Delay (QMD) is updated every 100 ms for packets sent in the last second. We also define Rate Error (RE) to evaluate traffic fading avoidance and use Jains fairness index [17] to assess fairness. The RE is calculated as:

$$\frac{SR_{actual} - SR_{allocated}}{SR_{allocated}} \quad (2)$$

Jains fairness index is normalized by the total throughput across all paths. We also measure video quality using SSIM [18] and MS-SSIM [19] over 2000 rendered frames.

A. Micro benchmark

We perform micro-benchmarks to assess throughput aggregation efficiency and the ability to avoid traffic fading.

Throughput Aggregations: In this scenario, we use one, two, and five subflows across three sub-scenarios. The paths mapped to subflows are disjoint, each with a 500 kbps bottleneck capacity. The test lasts for 120 s, and the subflows are joined at the start. The measurement summary is shown in Table II. FRACTaL starts conservatively, causing subflows to ramp up slowly. The fractional distribution algorithm manages the overall ramp-up, meaning that the media source using more subflows reaches the bottleneck capacity later than one with fewer subflows, resulting in a lower ABU in the short term for scenarios with more subflows.

[1]https://github.com/balazskreith/docker-gst-mprtp

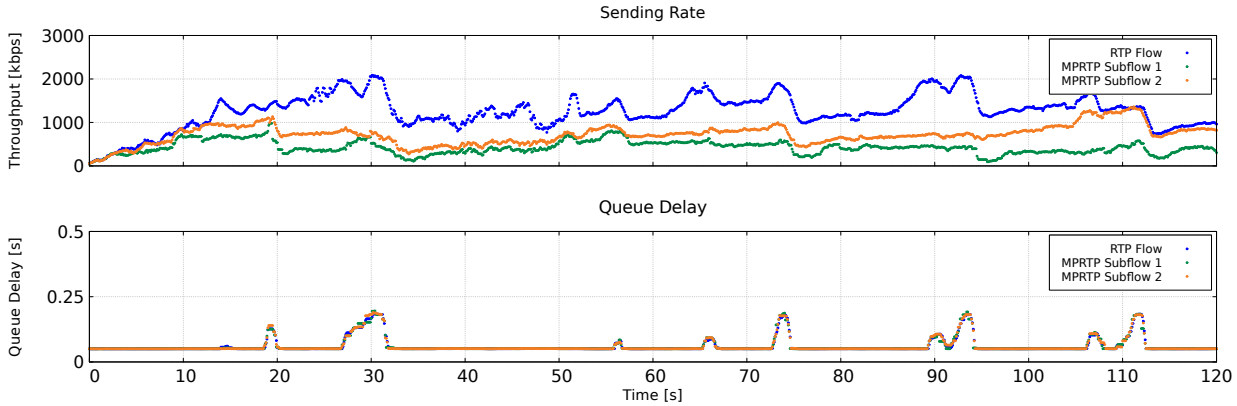


Fig. 5. The sending bitrate and the queue delay for a single- and a multipath capable media sources sharing the same bottleneck link.

TABLE II
SUMMARY OF THE MEASUREMENTS MEASURING THROUGHPUT AGGREGATIONS.

	1 Subflow	2 Subflows	5 Subflows
GP [kbps]	286 ± 17	515 ± 21	1205 ± 37
ABU [%]	76% ± 2%	67% ± 2%	60% ± 1%
LR [%]	0% ± 0%	0% ± 0%	0% ± 0%
QMD [ms]	107.81 ± 1.25	103.98 ± 0.54	102.66 ± 0.38

Avoiding Fading Effect: In this scenario, we evaluate the packet scheduler algorithm's ability to mitigate traffic fading. Two subscenarios are considered. In the first, one subflow uses a path with a bottleneck capacity varying between 500 kbps and 1000 kbps, with changes occurring every 25 s. In the second, two subflows are used, each with a bottleneck capacity varying between 500 kbps and 1000 kbps, but the aggregated capacity is 1500 kbps. The *Rate Error* (RE) is calculated for each subflow, and the results are summarized in Table III.
When the bottleneck capacity reduces on a path, FRACTaL detects congestion and recalculates the $SR_{estimated}$ for the subflow. The fractional distribution algorithm adjusts the $SR_{allocated}$ for each subflow. The MRCM requests a new $MR_{demanded}$ from the media source and forwards the updated $SR_{allocated}$ values to the MPPM. The MPPM then uses the packet scheduler algorithm to calculate the $SR_{transient}$, preventing traffic fading during the bitrate change. Consequently, the subflows' SR_{actual} fluctuates around the $SR_{transient}$, avoiding traffic fading.

TABLE III
SUMMARY OF MEASUREMENTS TEST FADING EFFECT BETWEEN SUBFLOWS

	1 Subflow	2 Subflows
GP [kbps]	369 ± 29	621 ± 28
RE [%]	16% ± 1%	16% ± 1%
LR [%]	1% ± 0%	0% ± 0%
QMD [ms]	110.2 ± 3.54	103.5 ± 0.74

Quality of Transferred Media Content: In this scenario, we compare the quality of the received media content with the original. We measure SSIM [18] and MS-SSIM [19] scores for the encoded media before sending and the decoded media after receiving. We use one, two, and five subflows, with disjoint paths and a total bottleneck capacity of 5000 kbps, divided equally among the subflows. SSIM and MS-SSIM are averaged over 2000 frames, with the average and standard deviation calculated over 30 measurements. The results, summarized in Table IV, show similar SSIM scores but better MS-SSIM scores when using multiple paths.

TABLE IV
SUMMARY OF MEASUREMENTS FOR SSIM AND MSSIM.

	1 subflow	2 subflows	5 subflows
SSIM	0.78 ± RUN	0.82 ± RUN	0.84 ± RUN
MSSIM	0.53 ± RUN	0.61 ± RUN	0.73 ± RUN

B. Fairness tests

We evaluate the fractional distribution algorithm's fairness by calculating Jain's fairness index for media sources using single and multiple paths. For a single path, fairness is the congestion control algorithm's ability to equally share the bottleneck capacity between flows. For multiple paths, fairness considers all flows across all paths.

Our test includes three scenarios: 1) A multipath source competes with a single-path source on a shared bottleneck; 2) A multipath source competes with a single-path source on a shared bottleneck and uses a non-shared one; 3) Two multipath sources compete over shared and non-shared bottleneck links.
One Bottleneck with Multiple Media Sources: We use two media sources on a shared bottleneck link, with one using two subflows and the other a single RTP flow. The shared bottleneck has a capacity of 3000 kbit/s (Figure 5). A summary of the measurements is shown in Table V.

Due to the fractional distribution algorithm, the multipath source ramps up more slowly, reaching saturation later than the single-path source. As a result, each subflow of the multipath source uses less bandwidth than the single flow. We calculated Jains index by normalizing throughputs with the available bandwidth of the used path. The Jains fairness index for this scenario is 0.97 ± 0.03. Due to FRACTaL's conservative nature around the bottleneck, its early congestion detection, and FEC protection during ramp-up, the flows do not experience packet loss.

TABLE V
SUMMARY OF MEASUREMENTS EVALUATING FAIRNESS BETWEEN A SINGLE- AND A MULTIPATH CAPABLE MEDIA SOURCE USING A SHARED BOTTLENECK.

	RTP flow	MPRTP flow 1	MPRTP flow 2
GP [kbps]	954 ± 180	464 ± 139	616 ± 113
ABU [%]	37% ± 6%	19% ± 5%	24% ± 4%
LR [%]	0% ± 0%	0% ± 0%	0% ± 0%
QMD [ms]	113.6 ± 4.36	113.62 ± 4.34	113.82 ± 4.15

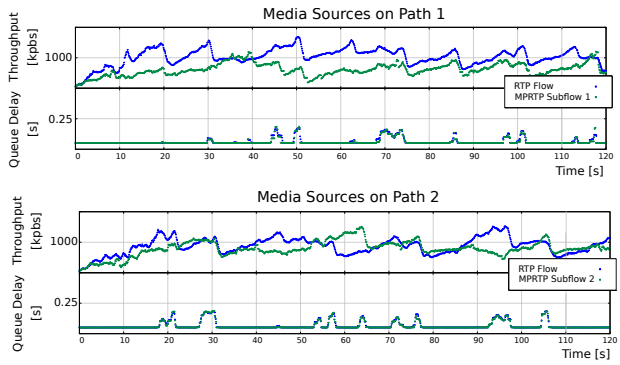


Fig. 6. The sending rate and the queue delay for a single- and a multipath capable media sources on different paths.

Two Bottlenecks with Two Media Sources: We use two media sources on a shared bottleneck link, with one using two subflows and the other a single RTP flow. Both paths have a capacity of 2000 kbit/s, and the test duration is 120 s. A summary of the measurements is shown in Table VI.

The multipath source ramps up slower than the single-flow source, leading to the single-flow source occupying a larger portion of the shared path. The Jains index is normalized by the aggregated bandwidth for multipath flows and by the single bandwidth capacity for single-path flows. The fairness index for this scenario is 0.98 ± 0.03.

TABLE VI
SUMMARY OF MEASUREMENTS EVALUATING FAIRNESS BETWEEN A SINGLE- AND A MULTIPATH CAPABLE MEDIA SOURCESS USING TWO DISJOINT PATHS.

	RTP flow	MPRTP flow 1	MPRTP flow 2
GP [kbps]	783 ± 161	428 ± 85	975 ± 401
ABU [%]	47% ± 8%	26% ± 4%	57% ± 20%
LR [%]	0% ± 0%	0% ± 0%	0% ± 0%
QMD [ms]	106.9 ± 2.75	106.2 ± 2.53	102.48 ± 1.95

Two Bottlenecks with Three Media Sources: We use three media sources on two shared bottleneck links, with one multipath capable source using two subflows across both paths, and one single-flow capable source on each path. Both paths have a capacity of 2000 kbit/s, and the test duration is 120 s (Figure 6). A summary of the measurements is shown in Table VII.

Due to the fractional distribution algorithm, the multipath source ramps up slower than the single-flow sources, resulting in a 2:1 bandwidth ratio in favor of the single-flow sources on each path. However, using aggregated throughput, the Jains index for this scenario is 0.95 ± 0.03.

TABLE VII
SUMMARY OF MEASUREMENTS EVALUATING FAIRNESS BETWEEN A SINGLE- AND A MULTIPATH CAPABLE MEDIA SOURCE USE DIFFERENT BOTTLENECK LINKS

	RTP flow 1	RTP flow 2	MPRTP flows
GP [kbps]	537 ± 151	487 ± 120	994 ± 100
LR [%]	0% ± 0%	0% ± 0%	0% ± 0%
QMD [ms]	106.92 ± 1.73	104.9 ± 1.5	105.91 ± 0.61

Two Bottlenecks with Two Multipath Capable Media Sources: We use two multipath capable media sources on two shared bottleneck links, each with a capacity of 2000 kbit/s, and the test duration is 120 s. The summary of the measurements is shown in Table VIII. The fractional distribution algorithm regulates both sources, ensuring they share bandwidth equally. The fairness index, calculated using aggregated throughput, is 0.99 ± 0.01.

TABLE VIII
SUMMARY OF MEASUREMENTS EVALUATING FAIRNESS BETWEEN TWO MULTIPATH CAPABLE SOURCES.

	MPRTP Flows 1	MPRTP Flows 2
GP [kbps]	565 ± 69	559 ± 98
LR [%]	0% ± 0%	0% ± 0%
QMD [ms]	106.31 ± 2.65	107.28 ± 1.89

C. Summary

This section evaluated the performance of our proposed algorithms: the packet scheduler and fractional distribution algorithms. We measured the packet scheduler's ability to avoid traffic fading and aggregate throughput. Fairness tests show that the fractional distribution algorithm ensures fairness among subflows across all paths. Our evaluation demonstrated that the packet scheduler algorithm successfully keeps the SR_{actual} around the $SR_{transient}$, avoiding traffic fading while aggregating path capacities. The fractional distribution algorithm regulates subflow ramp-up, leading to slower increases, but maintaining fairness for multiple paths and cross-traffic. No quality degradation was observed in video measurements using single- and multiple-paths.

VI. SYSTEMS CONSIDERATIONS

In this section, we detail a multimedia system implementing Multipath Rate Control using MPRTP [20]. The system sets up and tears down sessions (using RTSP or SIP), encodes/decodes media (e.g., VP8, VP9, H.265), applies SRTP encryption/decryption, and sends packets over multiple paths. It also supports DTLS, STUN, and ICE protocols for NAT traversal [21]. Figure 7 illustrates a component-based design for such a system based on our proposal.

The system has several components: Media Source/Sink (producing/consuming content), Media Encoder/Decoder (encoding/decoding RTP packets), packet sender/receiver, a dejitter buffer (for correct packet order), and a packet scheduler/rate controller (which maps RTP packets to subflows and controls bitrates).

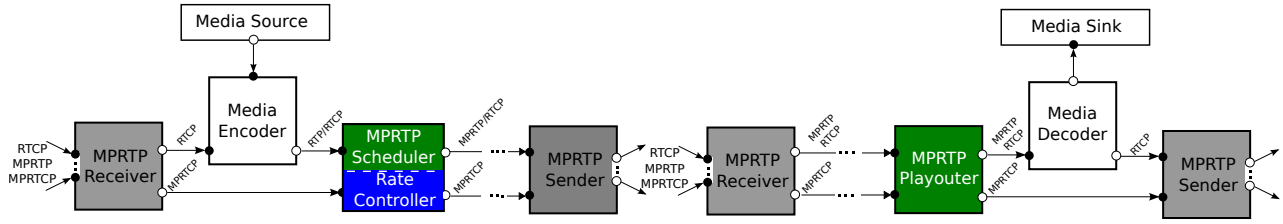


Fig. 7. High-level media pipeline design for a system implementing multipath rate control.

Both sender and receiver sides include a Multipath Sender and Receiver. The **Multipath Sender** transmits MPRTP/MPRTCP packets across mapped subflows, while non-MPRTP packets go through one path. The **Multipath Receiver** collects packets from all interfaces and outputs multiplexed MPRTP packets and non-MPRTP packets (e.g., MPRTCP, SDP, STUN).

On the sender side, the **Multipath Scheduler and Rate Controller** selects subflows for RTP packets and maps them to MPRTP packets. The Rate Controller monitors congestion and applies the fractional distribution algorithm to calculate the sending bitrate.

On the receiver side, the **Multipath Playouter** uses a dejitter buffer [8] to reorder packets for playback and sends MPRTCP packets for congestion control.

Compatibility

WebRTC applications use Secure RTP (SRTP) for encryption. The proposed subsystem enables RTP packets to operate over multiple paths without modifying other components. Since the Multipath Scheduler and Rate Controller alter RTP headers, security (e.g., encryption/decryption) must occur after header changes on the sender side and before playout on the receiver side.

VII. Discussion and Related Work

Much work has focused on congestion control for real-time media [4]–[6], [10], [22]–[25] and multipath capabilities for network protocols [26]–[29], but congestion control using multipath for real-time media is underexplored. Congestion control algorithms for real-time media impose strict delivery times. Many use packet loss as a congestion and bandwidth indicator [30], but retransmission is impractical as packets may miss the playout point. Monitoring queue delay improves reliability, but these algorithms are outperformed by loss-based methods. The IETF's RTP Media Congestion Avoidance Technique (RMCAT) working group developed congestion control algorithms considering queue delay fluctuations while competing with loss-based methods. NADA [6], [22] uses delay- and loss-based modes, estimating queuing delay via per-packet inter-arrival times. SCReAM [4] and Google Congestion Control (GCC) [5] model delay variation and apply filters for estimating queuing delay. FEC-based Rate Adaptation [23], [31] (FBRA) combines queue-delay congestion detection with error protection, while FRACTaL [10] improves FBRA for RMCAT. Multipath protocols like MPTCP [27], SCTP [32], and QUIC [29] use multiple paths for robustness. MPTCP design principles focus on cross-traffic fairness during bulk data transfer [9]. Recent MPTCP evaluations [33] highlight limitations in real-time applications, with cross-layer solutions for multipath communication in ad hoc networks [34] not optimized for real-time needs. For multimedia delivery, we used Multipath Real Time Protocol (MPRTP) [8], as MPTCP and SCTP dont consider real-time boundaries. MPRTP shifts traffic from congested to non-congested paths, though its path capacity exploration is limited. This research prompted further investigation into MPRTP with congestion control algorithms for all paths without media bitrate limitations. Optimized BBR [35] has shown promise in enhancing media delivery by optimizing bandwidth and reducing latency. Real-time media imposes strict time constraints, prohibiting bitrate throttling and retransmission. For interactive communication, media content may need rebalancing due to bandwidth limitations, with bitrate adjustments lagging behind production. Congestion control must ensure cross-traffic fairness [36]. Emerging research refines multipath congestion control, focusing on delay-based methods and machine learning to address real-time media delivery in heterogeneous networks [33], [36], [37], shaping future protocols and standards for real-time multipath communication.

VIII. Conclusions

This paper presented a rate control system for real-time media over multiple paths. The goal was to ensure fairness, avoid traffic fading, and efficiently utilize bandwidth. A congestion control algorithm was applied to all paths, with the ramp-up phase regulated for fairness. To avoid traffic fading, the transient sending bitrate and media source time lag were considered.

We proposed a layered architecture with the MRCM for congestion control and the MPPM for packet distribution. We developed algorithms for sending rate allocation and packet scheduling, implemented as a Gstreamer plugin. The system was tested in an emulated network environment, showing that our approach avoids traffic fading and maintains fairness.

We also discussed implementation details for multimedia systems and plan to apply the system in WebRTC, evaluating performance with RMCAT congestion control algorithms in future work.

REFERENCES

[1] N. Blum, S. Lachapelle, and H. Alvestrand, "Webrtc: Real-time communication for the open web platform," *Communications of the ACM*, vol. 64, no. 8, pp. 50–54, 2021.

[2] ITU-T, "G.114 One-way transmission time," *SERIES G: TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS International telephone connections and circuits General Recommendations on the transmission quality for an entire international telephone connection*, pp. 1–20, 2003.

[3] IETF. (2017) RMCAT working group of the IETF. [Online]. Available: http://datatracker.ietf.org/wg/rmcat

[4] I. Johansson and Z. Sarker, "Self-clocked rate adaptation for multimedia," 2015.

[5] H. Alvestrand, S. Holmer, and H. Lundin, "A Google Congestion Control Algorithm for RTCWEB," 2011, iETF Draft, draft-alvestrand-rtcweb-congestion.

[6] X. Zhu, R. Pan, S. Mena, P. Jones, J. Fu, S. D'Aronco, and C. Ganzhorn, "NADA: A unified congestion control scheme for real-time media," 2013.

[7] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," RFC 3550 (Internet Standard), RFC Editor, Fremont, CA, USA, pp. 1–104, Jul. 2003, updated by RFCs 5506, 5761, 6051, 6222, 7022, 7160, 7164, 8083, 8108. [Online]. Available: https://www.rfc-editor.org/rfc/rfc3550.txt

[8] V. Singh, S. Ahsan, and J. Ott, "Mprtp: multipath considerations for real-time media," in *Proceedings of the 4th ACM Multimedia Systems Conference*. ACM, 2013, pp. 190–201.

[9] D. Wischik, C. Raiciu, A. Greenhalgh, and M. Handley, "Design, Implementation and Evaluation of Congestion Control for Multipath TCP," in *Proc. USENIX NSDI*, 2011.

[10] B. Kreith, V. Singh, and J. Ott, "Fractal: Fec-based rate control for rtp," in *Proceedings of the 2017 ACM on Multimedia Conference. ACM*, 2017, pp. 1363–1371.

[11] gstreamer.org, "Gstreamer." [Online]. Available: http://gstreamer.org/

[12] openwebrtc.org, "OpenWebRTC." [Online]. Available: http://openwebrtc.org/

[13] S. Hemminger, "Network Emulation with NetEm," *Proceedings of the 6th Australian National Linux Conference (LCA 2005)*, no. April, pp. 1–9, 2005.

[14] Netem, "Traffic Controller for Netem." [Online]. Available: http://man7.org/linux/man-pages/man8/tc-netem.8.html

[15] Xiph.org, "YUV Sequences." [Online]. Available: http://media.xiph.org/video/derf/

[16] J. Bankoski, P. Wilkins, and Y. Xu, "Technical overview of VP8, an open source video codec for the web," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, IEEE. IEEE, 2011, pp. 1–6.

[17] R. Jain, A. Durresi, and G. Babic, "Throughput fairness index: An explanation," in *ATM Forum contribution*, vol. 99, no. 45, 1999.

[18] Z. Wang, "The ssim index for image quality assessment," *https://ece.uwaterloo.ca/~z70wang/research/ssim*, 2003.

[19] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, vol. 2. IEEE, 2003, pp. 1398–1402.

[20] V. Singh, T. Karkkainen, J. Ott, and S. Ahsan, "Multipath RTP (MPRTP)," 2012, iETF Draft, *draft-singh-avtcore-mprtp*.

[21] T. Lévai, B. E. Kreith, and G. Rétvári, "Supercharge webrtc: Accelerate turn services with ebpf/xdp," in *Proceedings of the 1st Workshop on eBPF and Kernel Extensions*, 2023, pp. 70–76.

[22] S. D'Aronco, L. Toni, S. Mena, X. Zhu, and P. Frossard, "Improved Utility-Based Congestion Control for Delay-Constrained Communication," *IEEE/ACM Transactions on Networking (TON)*, vol. 25, no. 1, pp. 349–362, 2017.

[23] M. Nagy, V. Singh, J. Ott, and L. Eggert, "Congestion Control using {FEC} for Conversational Multimedia Communication," *arXiv*, no. c, pp. 1–13, 2013. [Online]. Available: http://arxiv.org/abs/1310.1582

[24] M. Engelbart and J. Ott, "Congestion control for real-time media over quic," in *Proceedings of the 2021 Workshop on Evolution, Performance and Interoperability of QUIC*, ser. EPIQ '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 17. [Online]. Available: https://doi.org/10.1145/3488660.3493801

[25] A. S. Jagmagji, H. D. Zubaydi, S. Molnár, and M. Alzubaidi, "Utilizing machine learning as a prediction scheme for network performance metrics of self-clocked congestion control algorithm," *Infocommunications J.*, vol. 16, no. 3, pp. 2–17, 2024.

[26] J. R. Iyengar, P. D. Amer, and R. Stewart, "Concurrent multipath transfer using sctp multihoming over independent end-to-end paths," *IEEE/ACM Transactions on networking*, vol. 14, no. 5, pp. 951–964, 2006.

[27] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses," RFC 6824 (Experimental), RFC Editor, Fremont, CA, USA, pp. 1–64, Jan. 2013. [Online]. Available: https://www.rfc-editor.org/rfc/rfc6824.txt

[28] K. Rojviboonchai and H. Aida, "An evaluation of multi-path transmission control protocol (m/tcp) with robust acknowledgement schemes," *IEICE transactions on communications*, vol. 87, no. 9, pp. 2699–2707, 2004.

[29] Q. De Coninck and O. Bonaventure, "Multipath quic: Design and evaluation," in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*. ACM, 2017, pp. 160–166.

[30] A. Afanasyev, N. Tilley, P. Reiher, and L. Kleinrock, "Host-to-host congestion control for tcp," *IEEE Communications surveys & tutorials*, vol. 12, no. 3, pp. 304–342, 2010.

[31] J. O. Varun Singh Marcin Nagy, "Congestion Control Using FEC for Conversational Media," *draft-singh-rmcat-adaptive-fec*, 2013.

[32] S. Shailendra, R. Bhattacharjee, and S. K. Bose, "Mpsctp: A multipath variant of sctp and its performance comparison with other multipath protocols," in *2012 IEEE International Conference on Communication Systems (ICCS)*. IEEE, 2012, pp. 280–284.

[33] Ł. P. Łuczak, P. Ignaciuk, and M. Morawski, "Evaluating mptcp congestion control algorithms: Implications for streaming in open internet," *Future Internet*, vol. 15, no. 10, p. 328, 2023.

[34] M. Aljubayri, Z. Yang, and M. Shikh-Bahaei, "Cross-layer multipath congestion control, routing and scheduling design in ad hoc wireless networks," *IET Communications*, vol. 15, no. 8, pp. 1096–1108, 2021.

[35] S. Zhang and W. Lei, "An optimized bbr for multipath real time video streaming," *arXiv preprint arXiv:1901.09177*, 2019.

[36] M. Maliha, G. Habibi, and M. Atiquzzaman, "A survey on congestion control and scheduling for multipath tcp: Machine learning vs classical approaches," *arXiv preprint arXiv:2309.09372*, 2023.

[37] M. Pieska, A. Kassler, A. Brunstrom, V. Rakocevic, and M. Amend, "Performance impact of nested congestion control on transport-layer multipath tunneling," *Future Internet*, vol. 16, no. 7, p. 233, 2024.

Balázs Kreith is an experienced Software Engineer, focusing on Real-Time Communication Technologies (RTC). He graduated from the University of Debrecen, where he obtained his degree in Computer Science. Since 2016, he has been working in the field of RTC, specializing in developing solutions that enhance real-time communication performance, scalability, and reliability.

Árpád Drozdy graduated as an electrical engineer from the technical university of Budapest. He received his PhD. degree from Aalto University, Finland. He is specialized in 5G wireless telecommunications, holds three best paper awards, and currently works for Nokia.

Completion Time Prediction of Open Source FaaS Functions

David Balla, Markosz Maliosz, and Csaba Simon

***Abstract*—Function as a Service (FaaS) is the latest stage of application virtualization in the cloud. It enables to deploy small code pieces – functions – in the cloud. FaaS focuses on event-driven functions in response to triggers from different sources. The functions run in ephemeral virtual environments. This means that the user is charged on the basis of the time the function is busy serving the invocation requests. With the advent of Industry 4.0 the need has arisen to run applications on Edge Computing nodes. FaaS is a promising solution for serving industrial applications that require predictable latency while meeting the demands of edge computing, which operates on a limited resource base. Therefore, knowing the completion time of the invocation requests is of key importance.**

In this paper, we introduce a function runtime design for open-source FaaS implementations that achieves a lower deviation in request completion times compared to default runtimes by regulating the function's access to host CPU cores. We present the implementation details of our proposed function runtime design for Python, Go and Node.js. We also introduce a simulation framework that is able to estimate the completion time distribution of the incoming invocation requests. We validate the results of our simulation framework using real measurement data.

***Index Terms*—Communications Society, IEEE, IEEEtran, journal, LATEX, paper, template.**

I. INTRODUCTION

The granularity of application virtualization technologies is continuously increasing in the past decades, virtual machines have been partially replaced by light-weight container virtualization solutions, however, in both cases complete applications are hosted in the virtualized environment. In contrast to Virtual Machines (VMs) and containers, FaaS makes it available to host only a single building block of a distributed application in the virtualized environment. FaaS implementations are based on virtualization solutions, such as containers and lightweight VMs, such as microVMs [1], [2] or unikernels [3]. The FaaS framework spins up an instance for the incoming requests and keeps it running for a given time period and they are evicted if no requests are sent to them during this time. FaaS shows a change in the paradigm, a new way of application orchestration, rather than introducing a new implementation for virtualization. Functions can be organized into function chains, where one function invokes the next, thus implementing a complex application.

Functions are invoked for incoming events and they use the compute resources only during the execution. The corresponding billing model is based on the time the function keeps the compute resources allocated to serve the incoming requests. In contrast, using VMs and containers, the user pays for the up-time of the virtualized environment even if the hosted applications are idle.

The importance of FaaS can be seen by the fact that not only the major public cloud providers have implemented their FaaS platforms (e.g. Google Cloud, Amazon Web Services, Microsoft Azure, Alibaba Cloud), but the open source community has also embraced the technology. Several open-source FaaS implementations are available on GitHub, such as OpenFaaS [4], Fission [5], Kubeless [6] or Nuclio [7].

FaaS systems provide predefined function runtimes to execute the functions. Functions implemented by the users are encapsulated into the provided function runtimes, helping the users to focus on the business logic and sparing the manual integration work. However, these runtimes can majorly influence the performance of the functions.

Knowing the completion time of the requests has a key importance, as the billing model of FaaS systems is based on the combination of time the function is busy serving the requests as well as the allocated resources for the functions. Cloud providers can define different strategies for cost calculation. In AWS Lambda [8] and Microsoft Azure Functions [9], users are charged based on a per-millisecond rate, whereas users of Google Cloud Functions [10] and Alibaba Function Compute [11] are billed in 100-millisecond increments. The amount of allocated resources determines the price of each time increment. Therefore we exclusively focus on the completion time of the function invocations, as the cost of an invocation can be derived from the completion time and the amount of allocated resources.

In this paper we introduce a function runtime design that is able to provide stable completion times by regulating the function's access to the host CPU cores, thereby preventing bottlenecks and achieving higher QoS. We also introduce a method to estimate the completion time of function invocations. We present our results by using compute intensive functions implemented in Python, Go and Node.js.

We base our work on our previous paper [12] in which we have introduced our function runtime design as well as an algorithm that is able to estimate the completion time distribution of the function invocations. However, in [12] we have implemented our function runtime design exclusively in Python, and we designed our algorithm to estimate the completion time of functions that support parallel request processing. In this paper, we introduce the implementation details and

David Balla, Markosz Maliosz, Csaba Simon are with the High Speed Networks Laboratory Department of Telecommunications and Artificial Intelligence Budapest University of Technology and Economics, Budapest, Hungary. (E-mail: balla/maliosz/simon@tmit.bme.hu)

DOI: 10.36244/ICJ.2025.2.7

performance characteristics of our function runtime design for Python, Go and Node.js. We also introduce an additional algorithm that is designed to estimate the completion time distributions of invocation requests sent to function instances supporting sequential request processing. The introduced algorithms are taking into account the resource assignment strategies as well as the expected load that is sent to the examined function instance.

In this paper we focus on open source FaaS systems, as in this case we have access to the source code of the whole ecosystem in contrast to public cloud providers' implementations.

Open source FaaS systems are almost exclusively implemented on top of Kubernetes. Kubernetes is an open source cloud orchestration platform for containers. The basic architecture of an open source FaaS system is depicted in Fig. 1. The gateway component is the entry point of the FaaS system. It works as a proxy and directs the requests to the appropriate functions. Function instances run in containers that are encapsulated by Kubernetes pods. Function instances of the same type are hidden by a Kubernetes service. The gateway component addresses the service of the given function and forwards the requests to the actual function instances.

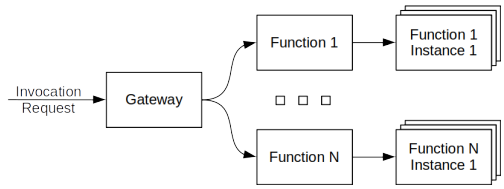


Fig. 1. Generic open source FaaS architecture

This paper is structured as follows. In Section II, we review the related work. Section III introduces our proposed function run-time design, while in Section IV we discuss the details of our function completion time prediction methods. In Sections V and VI we introduce our measurement environment and discuss our findings. Finally, Section VII concludes the paper.

II. Related Work

The benchmarking of FaaS implementations has a key importance, as different implementations have different performance characteristics [13], [14], [15]. M. Grambow et al. introduce BeFaaS [16], a framework to benchmark the major public providers' as well as open source FaaS platforms. BeFaaS comes with a built-in benchmark scenario that is an e-commerce application. BeFaaS can be extended by new benchmarks and can be created and added to the framework by using the BeFaaS programming library. In addition to showing the completion time of the functions, BeFaaS provides a drill-down analysis that helps to understand the particularities of the behavior of the benchmarked FaaS frameworks. P. Maissen et al. implemented FaaSdom, a tool to benchmark major public FaaS frameworks. FaaSdom [18] supports the benchmarking of different language runtimes, as well as it provides the ability to calculate the costs of the users. SeBS [17] provides a framework to benchmark public FaaS platforms. SeBS takes the following metrics into consideration, CPU and Memmory utilization, Response time, Code Size and Network IO. Costs can also be benchmarked by SeBS. The authors discuss the additional costs for the user that originates from the billing model of a particular FaaS provider.

Simulation frameworks can significantly reduce the costs of developing applications in a FaaS ecosystem, while also identifying factors that influence performance. SimFaaS [19] is a simulation framework for FaaS systems that enables the prediction of several performance metrics of a FaaS system, such as average response time, the probability of cold starts, and the average number of function instances. The authors validate the results of SimFaaS by comparing it to real usage data from AWS Lambda. M. Hanaforoosh et al. introduce MFS [20], a serverless FaaS simulator based on Apache Open Whisk. MFS calculates the reports of several metrics on top of the ones supported by SimFaaS, e.g., the number of functions that can or cannot be scheduled on any of the physical machines, the number of requests that can or cannot finish before a given deadline, and unlike SimFaaS, it handles containers and functions separately, reporting the number of containers used. SimLess [21] is a framework to simulate function choreographies in major public FaaS providers' ecosystems. SimLess considers various overheads such as network, concurrency and, authentication when estimating the round-trip time of the functions. J. Manner et al. introduces a methodology [22] to enable the comparison of the local and the cloud function execution and to map the local profiling data to the cloud platform. Their effort can significantly reduce development time as developers can work with their local tools that they are familiar with.

Using ephemeral, event-driven FaaS functions is a promising approach to implement services hosted on top of Edge computing devices, as such environments are equipped with a limited amount of compute resources. LambdaContSim [23] is a simulator designed to evaluate custom strategies to place functions on edge nodes to meet various requirements. It measures different metrics, such as placement success and failure, energy consumption, and service time. F. Filippini et al. present a simulation framework [24] for evaluating load-balancing algorithms in decentralized FaaS environments. The framework assesses performance using metrics such as success rate, power consumption, and the number of rejected requests.

Based on this review, our contribution differs from previous efforts, as our work focuses on the behavior of function runtimes under various compute resource assignment strategies, by using our proposed simulation framework.

III. Function Runtimes

Function runtimes are key components of FaaS systems. Function runtimes are encapsulating the user-defined functions, and by this lifting the burden of the integration of the functions to the distributed FaaS environment off the shoulders of the users. Function runtimes are implented as lightweight web-servers, that wrap the user-defined function. They also define an endpoint to respond to periodic health-check messages initiated by the FaaS framework. We have examined runtimes of several open source FaaS systems, namely

OpenFaaS [26], Fission [27], and Kubeless [28]. According to our investigations, the majority of the supported runtimes are capable of parallel request processing by starting a new worker thread for each of the incoming requests. However, these runtimes do not necessarily take into consideration the number of physical CPU cores that the host computer has. In case of starting more worker threads than the number of host CPU cores can lead to variable completion times. To overcome this issue, we propose a function runtime, that sets the limit the number of concurrently processed requests to the number of host CPU cores [12]. In the following, we introduce function runtimes for Python, Go and Node.js.

A. Python

Python runtimes are implemented by using different libraries in case of the examined FaaS frameworks. OpenFaaS and Fission implement their Python runtimes by using the Flask web framework, while Kubeless is using the Bottle library.

For the sake of simplicity, in the case of Python, we implemented our proposed runtime by using the Flask library. Flask supports parallel request processing by starting new Python threads or forking new Python processes. Python threads are sharing the Python interpreter's Global Interpreter Lock (GIL), which can lead to a serious performance bottleneck in the case of CPU intensive tasks [29]. To overcome this issue, we implemented our runtime by using Python's multiprocessing library, that starts a new Python interpreter with the user-defined function for each of the incoming requests. In Fig. 2, we show our measurement results performed on a computer having 16 physical CPU cores. We performed our measurements by sending function invocation requests to the examined function instance at different concurrency levels, which were integer multiples of the number of physical CPU cores. It can be seen that the tail latency is lower when limiting the number of worker threads to the number of available CPU cores.

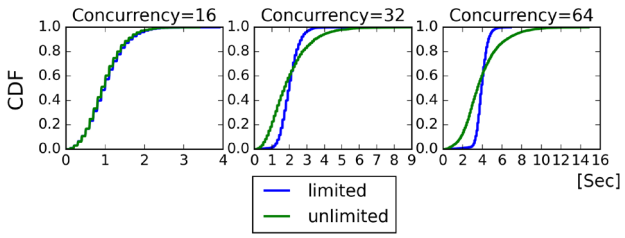


Fig. 2. Completion time differences - Limited vs. Unlimited number of worker threads - Python

B. Go

The Go runtimes of the examined open source FaaS implementations are all using the built-in HTTP server implemented in the Net library of the Go language that starts a new worker thread (Goroutine) for each of the incoming requests.

We extended the basic Go runtime design by adding a counting semaphore (implemented by Go buffered channels) to limit the number of simultaneously running worker threads. Fig. 3 depicts the performance difference of request processing between the proposed and the default function runtime designs. It can be seen that the proposed runtime design shows lower tail completion time values when the requests concurrency exceeds the number of CPU cores.

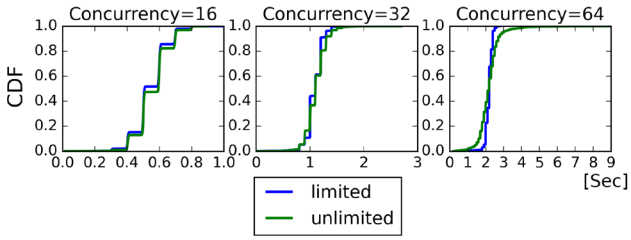


Fig. 3. Completion time differences - Limited vs. Unlimited number of worker threads - Go

C. Node.js

The Node.js runtimes are single threaded and are not capable of using multiple CPU cores simultaneously to concurrently serve CPU intensive tasks. By exploiting the Cluster or Worker threads library, we can achieve parallel processing of the incoming requests with Node.js. We have implemented a Node.js function runtime, that enables parallel request processing, by using the Worker threads library. To limit the number of concurrently running worker threads, we implemented our runtime to store all the incoming requests in a queue and use a counter to limit the number of requests processed concurrently. However, Node.js adds an extra layer of scheduling as it uses an event loop that allows to run tasks in an asynchronous way, therefore, the behavior of this function runtime differs from that runtimes where threads are directly scheduled by the operating system's task scheduler. Fig. 4 shows the results of using our proposed Node.js runtime with setting limited and unlimited number of worker threads to process the invocation requests.

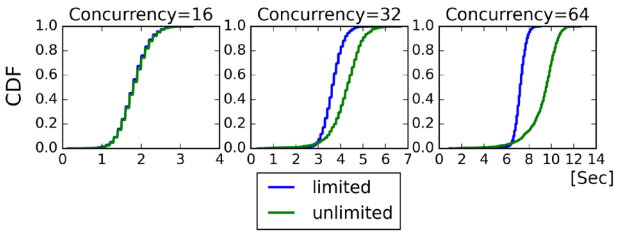


Fig. 4. Completion time differences - Limited vs. Unlimited number of worker threads - Node.js

IV. COMPLETION TIME PREDICTION

Being able to predict the completion time of function invocations is crucial. It makes available for the users to estimate their costs as the billing model of FaaS is based on the time the function spends on serving the incoming requests,

as well as it helps to design latency-sensitive workflows in a FaaS ecosystem.

In open source FaaS systems the function instances are running in containers encapsulated in Kubernetes Pods. Containerized applications are assigned to different Linux Control Groups (cgroups). Cgroups control the associated amount of resources to the application, e.g., CPU, memory, network. In open source FaaS frameworks, the users can specify the assigned amount of CPU and memory resources to the functions. These values are configured in the cgroups of the function instance's container.

The amount of CPU resources assigned to the container determines the amount of CPU time the applications in the container can consume in a single scheduling window. This value can be higher than 100% in the case of multi-core systems. For example, if the user assigns 200% CPU resources to the containerized application that has 4 threads and is running on a computer with 4 cores, then each of the threads can run in 50% of the scheduling window. The Completely Fair Scheduler (CFS) of Linux takes into account the CPU resources allocated for the applications through the cgroups.

We propose two algorithms to estimate the completion time distribution of the function invocations. Both algorithms model the CFS scheduler in a simplified way, by using round-robin scheduling. Cgroups are organized in a hierarchical way, which results in a hierarchical way of scheduling. However, the round-robin scheduling can be used, as the cgroup of the function container has no sub-cgroups, and the function container hosts only the user-defined function, that is, a small and simple single or multi-threaded application.

A. Multi-Threaded Runtimes

Our proposed algorithm for multi-threaded function runtimes [12] takes into account the number of CPU cores, the time (translated from CPU ticks) required for the function to run on the CPU to serve the invocation request (onCPU time), the assigned amount of CPU resources, and the concurrency of the incoming load. The on-CPU time is not necessarily equal to the response time, as the worker thread can be preempted while serving the request. The algorithm assumes that the function instance uses our proposed function runtime and that the user specifies the desired amount of CPU resources to be allocated to the function.

Our proposed function runtimes have a main thread that starts a new worker thread for each of the incoming requests until the number of worker threads reaches the CPU core limit. The proposed algorithm distributes the worker threads over the CPU cores. The algorithm runs the tasks in the scheduling window and decreases the tasks' onCPU time by the minimum time-slice as well as the CPU quota for the current scheduling window. If the CPU quota has decreased to less than the minimum time-slice, the algorithm starts a new scheduling window (see Algorithm 1).

According to our investigations, the minimum time-slice is 4 milliseconds, while the length of the scheduling window is 100ms. Our estimate for the time required to fork a new worker thread is 1 ms. If a task does not need the whole time-slice to finish, the rest of the time-slice can be used by the rest of the worker threads.

Algorithm 1: Completion time prediction for simultaneous request processing [12]

```
Function AddNewTask(startTime):
    actualWindow -= TimeOfFork
    tasks.append({timeOnCpu, startTime, endTime=-1})
    parallelTasks -= 1
    numTasks += 1

Function RunTask(task):
    task.timeOnCPU -= minTimeSlice
    actualWindow -= minTimeSlice

loops=1
timeFrame = 100ms
actualWindow = CPUThrottle
minTimeSlice = 4ms
timeOfFork = 1ms
numTasks = 0
parallelTasks = numCPUCores
tasks, finishedTasks = []
while len(FinishedTasks) != N do
    while actualWindow > minTimeSlice do
        task = tasks.GetNextTask()
        if task is mainTask then
            if numTasks < concurrency then
                // first batch of requests
                if parallelTasks < numCPUs then
                    AddNewTask(0)
                end
            else
                // at least one task of the first batch
                   has finished
                if newTasks > 0 and parallelTasks < numCPUs then
                    AddNewTask(finishedTasks[-newTasks].endTime+
                      ElapsedTimeTillNewRequest)
                    newTasks -= 1
                end
            end
        else
            RunTask(task)
            if task.timeOnCPU <= 0 then
                task.endTime = loops*timeFrame
                SaveTaskRuntime(task)
                finishedTasks.append(Task)
                tasks.Remove(Task)
                parallelTasks -= 1
                newTasks += 1
            end
        end
    end
    loops++
    actualWindow = cpuQuota
end
```

B. Single Threaded Runtimes

Some of the function runtimes only support sequential request processing, therefore, we propose another algorithm to estimate the completion time distribution of such functions. Initially, the algorithm starts tasks equal to the concurrency level and stores them in a list and starts to run the first task. The simulator decreases the on-CPU time and the CPU quota for the actual scheduling window. If the on-CPU time of the task is less than or equal to zero, then the simulator starts to run a new task as well as it adds a new task to the end of the task list. If the CPU quota reaches a value that is less than the minimum time slice, the simulator starts a new scheduling window. If the CPU quota is not zero but is less than the minimum time slice, it can be used in the next scheduling window. (see Algorithm 2)

Algorithm 2: Completion time prediction for sequential request processing

```
currTime = 0
for i in [1..Concurrency] do
    tasks.Append(Task(startTime=currTime))
end
actualWindow = CPUThrottle
task = tasks.popFirst()
while len(FinishedTasks) != N do
    while task.onCPU > 0 and actualWindow > 0 do
        task.onCPU -= minRuntime
        actualWindow -= minRuntime
        currTime += minRuntime
    end
    if task.onCPU <= 0 then
        FinishedTasks.append(currTime - task.startTime)
        task = tasks.popFirst()
        tasks.Append(Task(startTime = currTime))
    end
    if actualWindow <= min_runtime then
        if actualWindow > 0 then
            actualWindow = CPUThrottle + actualWindow
        else
            actualWindow = CPUThrottle
            currTime += schedWindow - CPUThrottle
        end
    end
end
```

V. TESTBED

A. Test environment

We deployed our measurement environments by installing a two-node Kubernetes cluster on Cloudlab [25] c220g1 compute nodes equipped with 2 Intel E5-2630 v3 8-core and 128 GB RAM and 10 Gb network interfaces. The Linux CFS operates with time values instead of CPU ticks. However, if the CPU frequency scaling is turned on, the number of CPU ticks and the number of operations under a given time period are undefined. Therefore, we turned off the CPU frequency scaling and locked the CPU frequency to 2.2GHz. We have also turned off the logical CPU cores given by the HyperThreading capability, to avoid situations when tasks are scheduled to the same physical core.

B. Test Functions

We implemented our test functions in Python, Go, and Node.js programming languages. Each of the functions calculates the estimated value of π by using the Leibniz formula. The Leibniz formula takes a parameter that sets the accuracy of the estimated value of π. This parameter is the input value of the function. The higher the value of the parameter, the more accurate the estimated value of π. For our measurements, we selected the input values from a normal distribution.

C. Load Generator

For our measurements, we used Hey, an HTTP load generator that is able to generate HTTP requests with a given concurrency level. To maintain a stable concurrency level, Hey initially starts client threads equal to the desired concurrency level. First, all the clients send out their requests. After that, a client can only send a new request once a response has been received for the current request.

Hey is only able to work with the same request content during load generation. Therefore, we modified the codebase of Hey to be able to send different input values to the functions.

VI. EVALUATION

We have implemented the proposed algorithms as function completion time simulator software modules. We validate the accuracy of the proposed simulators by comparing their outputs with real measurement results. We performed our measurements by using function runtimes that support parallel and sequential request processing.

To show the ability of our simulator software to predict the completion time distribution even in the case of variable input, we performed our measurements by selecting the input variables from a normal distribution. In this case, the configuration of the simulator requires two measurements. To be able to calculate the values of a normal distribution, we need to know the mean and deviation values. Thus, we performed two measurements using the median and deviation of the input values, to get the corresponding onCPU times.

The onCPU times can be acquired by using the Linux Perf tool. However, running Perf requires system administrator rights that are not necessarily available in all cases. In such situations, we can estimate the onCPU time by the response time of the function invocation as shown by eq. (1). When using eq. (1) to estimate the onCPU value, we suppose that the user-defined function does not start any further threads. However, this calculated value includes the additional latency of the network transport.

$$OnCPU \approx \lfloor \frac{Comp.T}{Sch.W} \rfloor * CPU_{fn} + Comp.T - \lfloor \frac{Comp.T}{Sch.W} \rfloor * Sch.W \quad (1)$$

Where:

- $Sch.W$ = 100ms, scheduling window,
- $CPU_{fn} \in [1..100]$, amount of CPU resources assigned to the function,
- $Comp.T$: completion time of a single function invocation.

A. Parallel Request Processing

To show the efficiency of our simulator, we present our results of different scenarios covering cases with different CPU and concurrency settings. The performance of the examined language runtimes are different, therefore, we adjusted the function invocation parameters accordingly.

Fig. 5 shows our measured and simulated results in the case of runtimes that are able to process multiple requests simultaneously. It can be seen that the measured and simulated values are very close to each other. In Table I we summarized the absolute differences between the measured and simulated results, for each of the language runtimes, related to the median and the 95th percentile.

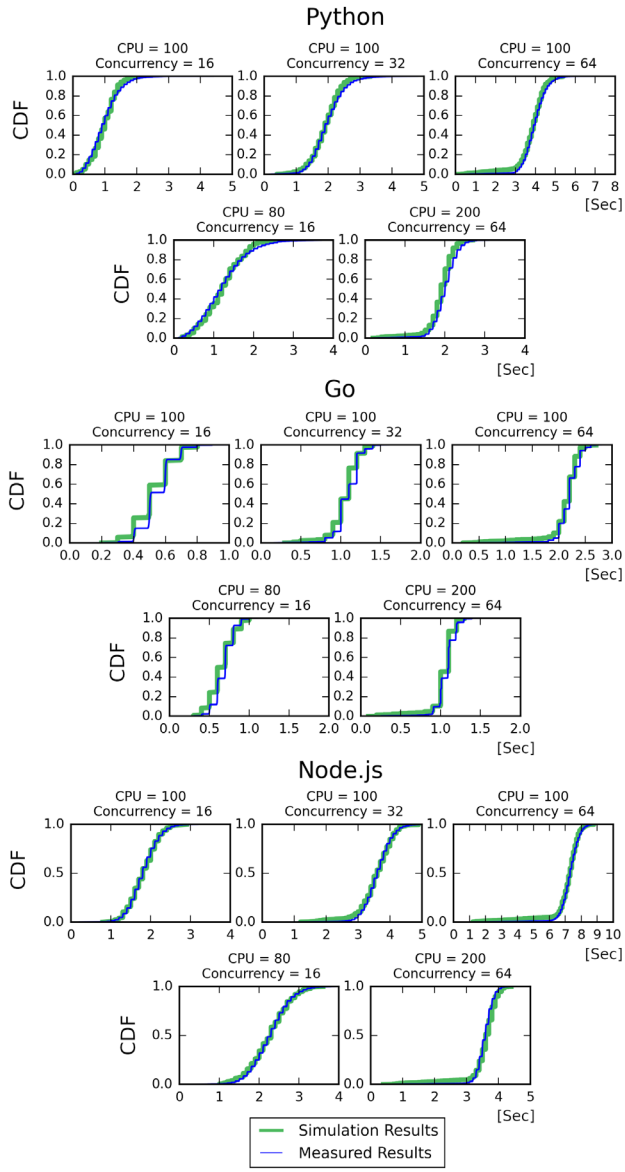


Fig. 5. Real vs Simulation completion time distribution in case of parallel request processing

TABLE I
MEDIAN AND 95^{th} %ILE ERRORS [SEC]
PARALLEL REQUEST PROCESSING

CPU	Conc.	Percentile	Python	Go	Node.js
80%	16	50th	0.0047	0.0038	0.0038
		95th	0.1974	0.0023	0.0078
100%	16	50th	0.0056	0.0081	0.0002
		95th	0.1987	0.0000	0.0022
	32	50th	0.0949	0.0999	0.0007
		95th	0.1007	0.0003	0.0008
	64	50th	0.0089	0.0000	0.0002
		95th	0.1953	0.0998	0.0014
200%	64	50th	0.0998	0.0003	0.1046
		95th	0.1907	0.0101	0.1104

B. *Sequential Request Processing*

Fig. 6 shows the results of our measured and simulated completion time distributions for each of the examined function runtimes. In this case, we selected the input values to generate jobs of which onCPU times are less than the time the function is scheduled to run in a given scheduling window. This leads to a scenario where some of the function invocation completion times are significantly longer than the rest, as it can happen that the function is preempted by the scheduler while it is processing an invocation request. In this case, not only the completion time of the preempted request is influenced, but all the completion times of the requests that arrived after the preempted request. We experienced an undefined behavior in the case of Node.js over the request concurrency level of 32, therefore, in Fig. 6 we only show our results for the concurrency levels of 16 and 32. We suppose that the experienced undefined behavior over the concurrency level of 32 is due to the asynchronous event-loop of the Node.js runtime. We show the absolute differences between the measured and simulation results related to the median and 95^{th} percentile in Table II.

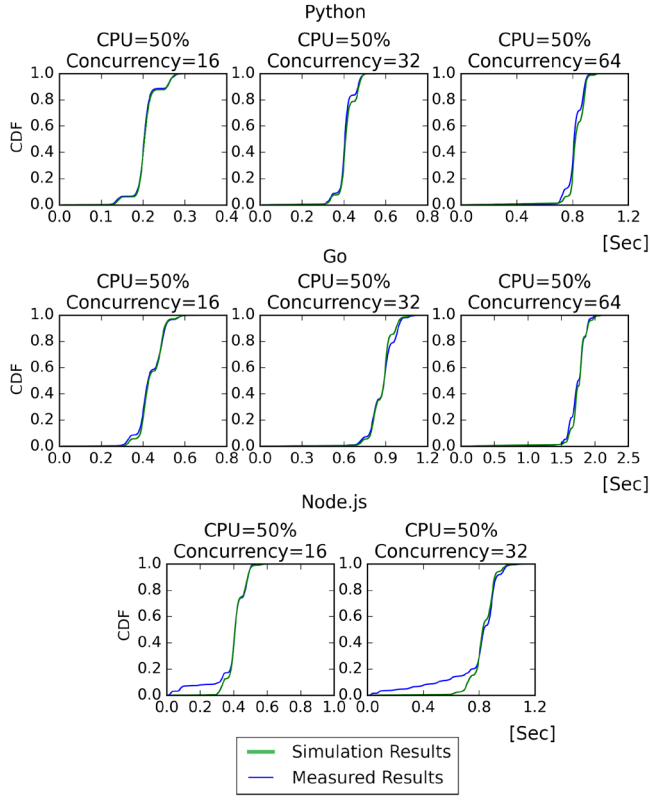


Fig. 6. Real vs Simulation completion time distribution in case of sequential request processing

C. *Differences of Sequential and Parallel Request Processing*

We investigated the completion time differences in the cases where function runtimes supporting parallel request processing

TABLE II
MEDIAN AND 95^{th} %ILE ERRORS [SEC]
SEQUENTIAL REQUEST PROCESSING

CPU	Conc.	Percentile	Python	Go	Node.js
50%	16	50th	0.0005	0.0079	0.0031
		95th	0.0022	0.0083	0.0052
	32	50th	0.0010	0.0015	0.0172
		95th	0.0019	0.0156	0.0384
	64	50th	0.0039	0.0240	-
		95th	0.0070	0.0015	-

and those where sequential processing runtimes were used. In Fig. 7 we show the differences between the completion time distributions related to the examined cases. In this scenario, we used our Python and Go language runtimes and assigned 100% of CPU resources to them. It can be seen that, in the case of Python, using a function runtime that supports simultaneous request processing results in higher completion time values in general. This phenomenon can be explained by the additional time that is caused by the starting a new Python interpreter for each of the incoming requests. In case of Go, the completion time distributions for the two cases are very close to each other. Also, our parallel processing Go runtime starts threads that are not as expensive as starting new processes. Parallel function runtimes can gain extra performance with more than 100% of assigned CPU resources.

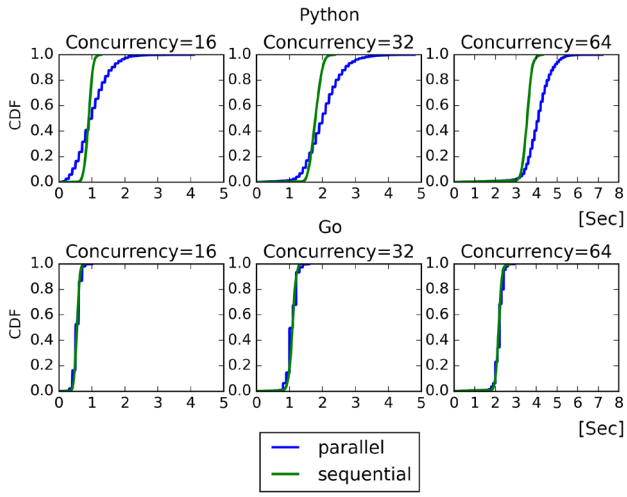


Fig. 7. Distribution of function completion times in case of parallel and sequential request processing runtimes

D. Variable CPU frequency

We performed our measurements by turning off the CPU frequency scaler. However, in a real-life scenario, the lack of enabling the CPU frequency scaler results in high energy consumption even if the compute node is idle. Therefore, we investigated the effects of the CPU frequency scaling on the accuracy of our measurements. We performed our measurements with the CPU frequency scaling turned on and off, as well as investigated the effects of CPU utilization generated by other CPU heavy loads. For our measurements, we used our function implemented in Go, with 50% of CPU assigned, to also have an idle period in the scheduling window. In this idle period the frequency scaler will scale down the CPU frequency short after the application is scheduled out even if the application remains on the same CPU core. To mitigate any unpredictable behavior on the software side, we generated load using a fixed input value instead of varying the input parameters. Fig. 8 shows the results of our measurements. It can be seen that the results are influenced by the variable CPU frequency caused by the CPU frequency scaler. In the case of a fully utilized compute node, the CPU clock runs on the maximal frequency, which results in more predictable completion times that are very close to the results of the scenario where the CPU frequency scaling is turned off. We can also observe that the background load does not influence the performance of the function. This can be explained by the assigned amount of CPU resources that is guaranteed for the function instance at the level of the scheduler of the OS.

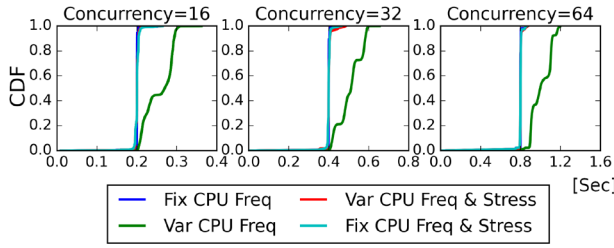


Fig. 8. Effects of CPU frequency scaling on function completion times

VII. CONCLUSION

In this paper we introduced a function runtime design that is able to process multiple requests at the same time but it limits the number of simultaneously processed requests to the number of available CPU cores. We showed that using our runtime design the tail completion times can be reduced significantly. We implemented the proposed function runtime design for Python, Go and Node.js.

We have also proposed an algorithm to predict the completion time distribution of the invocation requests sent to a FaaS function that are using our proposed function runtimes. However, not all the function runtimes support simultaneous request processing. Therefore, we proposed another algorithm that is able to forecast the completion time distribution of invocation requests sent to functions that are only capable of sequential request processing. We implemented both proposed algorithms as function runtime simulators and validated their results using real measurement data from scenarios with varying inbound request concurrency and CPU resource allocations for the functions.

Knowing the function completion time in advance can highly facilitate the design phase of software architectures based on FaaS considering the billing model of FaaS. FaaS operates with ephemeral function instances; therefore, FaaS is a promising model to implement services hosted by edge computing nodes. Knowing the completion times for a given load, our work helps engineers to better utilize the limited resource base of edge computing devices.

REFERENCES

[1] Agache, A., Brooker, M., Iordache, A., Liguori, A., Neugebauer, R., Piwonka, P., & Popa, D. M. (2020). Firecracker: Lightweight virtualization for serverless applications. In *17th USENIX symposium on networked systems design and implementation (NSDI 20)* (pp. 419–434).

[2] Randazzo, A., & Tinnirello, I. (2019, October). Kata containers: An emerging architecture for enabling mec services in fast and secure way. In *2019 Sixth International Conference on Internet of Things: Systems, Management and Security (IOTSMS)* (pp. 209–214). IEEE.

[3] Manco, F., Lupu, C., Schmidt, F., Mendes, J., Kuenzer, S., Sati, S., ... & Huici, F. (2017, October). My VM is Lighter (and Safer) than your Container. In *Proceedings of the 26th Symposium on Operating Systems Principles* (pp. 218–233).

[4] https://openfaas.com

[5] https://fission.io/

[6] https://github.com/vmware-archive/kubeless

[7] https://nuclio.io/

[8] AWS Lambda pricing, https://aws.amazon.com/lambda/pricing

[9] Azure Functions pricing, https://azure.microsoft.com/pricing/details/functions/

[10] Pricing Overview, https://cloud.google.com/functions/pricing-overview

[11] Function Compute Pricing, https://www.alibabacloud.com/product/function-compute/pricing

[12] Balla, D., Maliosz, M., & Simon, C. (2021, November). Estimating function completion time distribution in open source FaaS. In *2021 IEEE 10th International Conference on Cloud Networking (CloudNet)* (pp. 65–71). IEEE.

[13] Motta, M. A. C., Carvalho, L. R., Rosa, M. J. F., & Araujo, A. P. F. (2022). Comparison of faas platform performance in private clouds. *Proceedings of the 12th CLOSER*, 109–120.

[14] Ma, P., Shi, P., & Yi, G. (2023, October). Feature and Performance Comparison of FaaS Platforms. In *2023 IEEE 14th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 239–243). IEEE.

[15] Barcelona-Pons, D., & García-López, P. (2021). Benchmarking parallelism in FaaS platforms. *Future Generation Computer Systems*, 124, 268–284.

[16] Grambow, M., Pfandzelter, T., Burchard, L., Schubert, C., Zhao, M., & Bermbach, D. (2021, October). Befaas: An application-centric benchmarking framework for faas platforms. In *2021 IEEE International Conference on Cloud Engineering (IC2E)* (pp. 1–8). IEEE.

[17] Copik, M., Kwasniewski, G., Besta, M., Podstawski, M., & Hoefler, T. (2021, November). Sebs: A serverless benchmark suite for function-as-a-service computing. In *Proceedings of the 22nd International Middleware Conference* (pp. 64–78).

[18] Maissen, P., Felber, P., Kropf, P., & Schiavoni, V. (2020, July). Faasdom: A benchmark suite for serverless computing. In *Proceedings of the 14th ACM international conference on distributed and event-based systems* (pp. 73–84).

[19] Mahmoudi, N., & Khazaei, H. (2021). Simfaas: A performance simulator for serverless computing platforms. *arXiv preprint arXiv:2102.08904*.

[20] Hanaforoosh, M., Ashtiani, M., & Azgomi, M. A. (2023, May). MFS: A serverless FaaS simulator. In *2023 9th International Conference on Web Research (ICWR)* (pp. 81–86). IEEE.

[21] Ristov, S., Hautz, M., Hollaus, C., & Prodan, R. (2022, November). SimLess: simulate serverless workflows and their twins and siblings in federated FaaS. In *Proceedings of the 13th Symposium on Cloud Computing* (pp. 323–339).

[22] Manner, J., Endreß, M., Böhm, S., & Wirtz, G. (2021, September). Optimizing cloud function configuration via local simulations. In *2021 IEEE 14th International Conference on Cloud Computing (CLOUD)* (pp. 168–178). IEEE.

[23] Matricardi, A., Bocci, A., Forti, S., & Brogi, A. (2023,August). Simulating FaaS Orchestrations In The Cloud-Edge Continuum. In *Proceedings of the 3rd Workshop on Flexible Resource and Application Management on the Edge* (pp. 19–26).

[24] Filippini, F., Calmi, N., Cavenaghi, L., Petriglia, E., Savi, M., & Ciavotta, M. (2024, June). Analysis and Evaluation of Load Management Strategies in a Decentralized FaaS Environment: A Simulation-Based Framework. In *Proceedings of the 1st Workshop on Serverless at the Edge* (pp. 1–8).

[25] Duplyakin, D., Ricci, R., Maricq, A., Wong, G., Duerig, J., Eide, E., ... & Mishra, P. (2019). The design and operation of CloudLab. In *2019 USENIX annual technical conference (USENIX ATC 19)* (pp. 1–14).

[26] https://github.com/openfaas/templates

[27] https://github.com/fission/environments

[28] https://github.com/vmware-archive/runtimes

[29] Python GlobalInterpreterLock https://wiki.python.org/moin/GlobalInterpreterLock

David Balla is a PhD candidate at the Budapest University of Technology and Economics (BME), Department of Telecommunications and Artificial Intelligence, High Speed Networks Laboratory. His research interests include low-latency and real-time networking and cloud computing solutions. His current research focuses on cloud-native software architectures and real-time FaaS systems.

Markosz Maliosz is an associate professor at the Budapest University of Technology and Economics (BME), Department of Telecommunications and Artificial Intelligence, High Speed Networks Laboratory. He received his PhD (2006) and MSc (1998) degrees in Computer Science in the field of communication systems from BME. He worked as guest researcher and consultant at telecommunication in the areas of network performance evaluation and time-sensitive networks. His research interests include network architectures and design, optimization techniques and traffic engineering, his current research activity focuses on network virtualization and optimization, and cloud networking. He is member of the Scientific Association for Infocommunications (HTE), Hungary.

Csaba Simon obtained his PhD degree at Budapest University of Technology and Economics, Department of Telecommunications and Artificial Intelligence and he is working at the same Department since 2001. His research interests are mostly related to 5G and 6G Systems, virtualization, cloud native applications and network and service management.

Framework for Intrusion Detection in IoT Networks: Dataset Design and Machine Learning Analysis

Mansour Lmkaiti, Ibtissam Larhlimi, Maryem Lachgar, Houda Moudni and Hicham Mouncif

Abstract—**This study explores the development of robust Intrusion Detection Systems (IDS) to enhance cybersecurity in Wireless Sensor Networks (WSNs) within the evolving Internet of Things (IoT) ecosystem. It leverages a publicly available dataset derived from UNSW-NB15, retrieved from a GitHub repository, capturing diverse network traffic attributes (dttl, swin, dwin, tcprtt, synack, ackdat), protocol-specific indicators (proto tcp, proto udp), and service-specific attributes (service dns). These features enable precise analysis of TCP/IP headers and traffic patterns, supporting multi-class classification into four categories: Analysis, Denial of Service (DoS), Exploits, and Normal. Advanced machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), were applied with systematic preprocessing (including KNN-based imputation, normalization, and one-hot encoding), feature selection using Random Forest importance, and 5-fold cross-validation. The best performance was achieved by Random Forest (accuracy, precision, recall, and F1-score of 99.9877%), followed by KNN (99.9754%) and SVM (99.9630%). The study demonstrates that combining well-structured models with relevant protocol-level features and robust evaluation strategies can significantly enhance intrusion detection capabilities in IoT-based environments. It reinforces the value of using modern public datasets and interpretable algorithms for building scalable and reliable IDS solutions.**

Index Terms—**Artificial Intelligence (AI), Intrusion Detection Systems (IDS), Machine Learning (ML), Internet of Things (IoT),Wireless Sensor Networks (WSN)**

I. INTRODUCTION

In today's technological environments, where businesses, governments, and individuals must contend with constantly changing and complex threats, cybersecurity has emerged as a key component. Although the Internet of Things' (IoT) explosive growth has greatly improved connectivity and operational effectiveness, it has also made networked systems more complex and vulnerable. Wireless Sensor Networks (WSNs), as key components of IoT ecosystems, are particularly sensitive to multiple security concerns due to their resource-constrained nature, poor computing capacity, and frequent deployment in hostile and insecure situations. [1] gave a thorough rundown of how artificial intelligence (AI) may be included into sensor networks, highlighting how it can be used to mitigate these weaknesses and enhance network performance and scalability. In order to protect these networks from online attacks, intrusion detection systems, or IDS, are essential. IDS models have been benchmarked using traditional datasets such as KDD Cup 1999 and NSL-KDD, although it is well known that these datasets do not accurately represent contemporary network circumstances. [2] emphasized the importance of creating more advanced and diverse datasets to overcome these challenges, introducing features that better capture the complexities of current network dynamics. These datasets are crucial tools for developing machine learning models that can identify, evaluate, and eliminate risks instantly. An comprehensive dataset created especially for analyzing modern cyberattacks is presented in this article. Important features that provide thorough insights into TCP/IP are 'dttl', 'swin', 'dwin', and 'tcprtt'. Additionally, protocol-specific attributes (proto_tcp', proto_udp') and service-specific details ('service_dns') enable the recognition of patterns in network traffic. According to [3], multi-class classification – a critical element of trustworthy intrusion detection – is enhanced by grouping data intro groups such as Analysis, Exploits, Denail of Service (DoS), and Normal. This work builds on this dataset using advanced machine learning techniques including Random Forest, Support Vector Machines, and K-Nearest Neighbors. These algorithms have proven to be effective in handling complex network datasets and achieving high levels of intrusion detection precision. [4] demonstrated Random Forest exceptional performance in a range of intrusion detection scenarios, demonstrating its dependability and low mistake rates. Similarly, [5] demonstrated the potential of hybrid models, such as combining boosting techniques with KNN, to improve classification efficiency in diverse scenarios. Previous research has demonstrated the efficacy of integrating ML techniques into IoT and WSNs to address critical challenges such as energy efficiency, routing optimization, and realtime intrusion detection. For instance, [7] explored feature selection techniques to enhance both security and system performance. Additionally, [6] demonstrated how, especially in resource-constrained WSNs, hybrid technqiues like particle swarm optimization (PSO) in conjunction with neural networks and reinforcement learning may greatly improve intrusion detection capabilities. To sum up, this study offers a solid foundation for creating an IDS that is suited to the complexity of contemporary networks. As the Internet of Things and Wireless Sensor Networks expand quickly, new types of cyberthreats are appearing that are unable to counter

M. Lmkaiti, I. Larhlimi, M. Lachagr, Hicham Mouncif are affiliated with laboratory of Innovation and Mathematics, Applications and Information Technology, Polydisciplinary Faculty, Sultan Moulay Slimane University, Beni Mellal, Morocco.

(E-mails: lamkaitimansour@gmail.com, ibtissam.larhlimifpb@usms.ac.ma, maryamlachgar96@gmail.com, h.moudni@usms.ma, h.mouncif@usms.ma)

Manuscript received April 19, 2005; revised August 26, 2015.

DOI: 10.36244/ICJ.2025.2.8

with traditional network security solutions.Intrusion Detection Systems powered by machine learning have the potential to ovvercome these obstacles by identifying intricate patterns of malevolent activity. In this study, we leverage a publicly available dataset derived from UNSW-NB15 [22], [23] and hosted on GitHub, specifically structured to reflect contemporary IoT traffic and attack scenarios. This dataset includes protocol-level and service-specific attributes that are rarely present in older benchmarks such as KDD99 or NSL-KDD. Our primary objective is to conduct a detailed comparative analysis of three well-known ML algorithms [8] [10]: Random Forest, SVM, and KNN—within a multi-class classification framework adapted to modern IoT network conditions. Few recent IDS papers explore low-level protocol state features due to their parsing complexity and variability, this work addresses this gap. The proposed framework is designed to guide the development of scalable and reliable IDS solutions, particularly in smart city infrastructure and industrial IoT applications where detecting real-time threats is essential to maintaining service availability and data protection. By leveraging a rich dataset and advanced ML techniques, this study addresses critical cybersecurity challenges, paving the way for improved resilience in IoT and WSN environments. Unlike conventional approaches focused on flow statistics or paylod metadata, our method utilizies underexplored protocol-level features, such as TCP flags, service identifiers, and session behaviors. This allow for a unique balance between model interpretability and fine grained detection, which is often lacking in recent deep learning of flow based IDS studies.

II. Related Work

Large-scale traffic captures with contemporary attack simulations are provided by datasets such as IDSAI and BoTIoT; nonetheless, their generalizability may be constrained by severe class imbalance or excessively artificial behaviors. The dataset employed in this study, on the other hand, combines controlled labeling with true protocol-level variables to strike a compromise between realism and diversity. Unlike widely used datasets as IDSAI or BoT-IoT, which often suffer from severe class imbalance or overly synthetic behavior, the dataset employed in this study offers a more realistic and balanced distribution of traffic of traffic types. Its clearly labeled categories and protocol-level granularity make it a more robust foundation for evaluating machine learning models under pratical IoT conditions. It is more suitable fore fine-grained model evaluationin IoT-focused contexts due to its multiclass structure and clear attribute descriptions. Much recent research has focused on applying machine learning approaches to IoT networks, particularly in the areas of routing, energy efficiency, and security improvement. This section highlights significant contributions in various domains. El Khediri et al. [1] gave a thorough rundown of integrating AI into sensor networks, tackling issues like scalability and energy limitations that are essential for optimizing IoT networks. Their research highlighted Ai capacity to overcome resource constraints.

Gutierrez-Portela et al. [2] presented a new dataset (IDSAI) and illustrated how machine learning models may be used to detect intrusions in Internet of Things communications. Their research made clear how crucial reliable datasets are to enhancing ML model's flexibility in chnaging contexts.

Vanitha et al. [3] suggested a Bayesian machine learning method for WSN route optimization, emphazing effective deviation management and route selection. The importance of probabilistic models in enhancing network dependability was highlighted by this sudy.

Dharini et al. [4] investigated and proved the efficacy of boosting algorithms against DoS assaults in the context of intrusion detection in WSNs. Their results reaffirmed how ensemble learning approaches may be used to improve newtork security.

Suresh et al. [5] created a clever routing plan for IoT-enabled WSNs by utilizing deep reinforcement learnin. Their findings demonstrated notable gains in data transmission dependability and energy economy, making it a viable strategy for dynamic IoT contexts.

Yadav et al. [7] centered on methods for feature selection and classification to enhance IoT application security and performance. Their research showed that customized feature engineering greatly improves networks performance.

Narayanan et al. [6] enhanced intrusion detection systems by combining artificial neural networks with particle swarm optimization. Their hybrid strategy demonstrated how crucial it is to combine ML models and optimization techniques in order to adress difficult issues in IoT networks.

Surenther et al. [8] suggested a grouping model strategy made possible by machine learning to maximize energy use and data transfer effectiveness in WSNs. The necessity of energy-aware techniques in IoT installations was highlighted bu this study.

Tabbassum et al. [9] created a succesful fuzzy-based clustering algorithm for data transmission in WSNs, demonstrating how it can prolong network lifetime while preserving energy efficiency.

Lai et al. [10] used online learning methods to identify DoS assaults in WSNs, providing a scalable way to mitigate threats in real time. Similarly, Ayuba et al. [11] utilized ensemble ML models to enhance DoS detection, demonstrating the importance of adaptive frameworks in securing IoT networks. Recent developments in IoT security have emphasized the need for more adaptive and intelligent intrusion detection mechanisms. Studies have explored federated learning [19] to enhance privacy and decentralization in IDS architectures [18], [19], especially in industrial environments where data sharing is sensitive. Edge computing is also gaining traction, enabling real-time detection with reduced latency by processing data closer to the source. Additionally, several works focus on zeroday attack [17] detection by leveraging behavior-based models capable of identifying previously unseen threats. Contextaware intrusion detection systems [20] have also emerged as promising approaches, adapting their detection logic to the operating environment of IoT devices. Despite these advances, many of these solutions remain fragmented, and few

combine these techniques into a unified, scalable framework. Our work addresses this gap by focusing on protocol-level features while proposing a flexible model structure that can be integrated with such advanced strategies in future iterations [21]. In recent years, several deep learning techniques such as Convolutional Neural Networks (CNN) [24], Long Short-Term Memory (LSTM) [24] networks, and transformer-based models have been applied to intrusion detection in IoT and WSN environments. These methods are particularly effective at identifying hidden and sequential patterns in network traffic. In addition, they often require large-scale labeled datasets and significant computational resources, which limits their applicability in real-world, resource-constrained systems. Moreover, some studies have explored adaptive or online IDS frameworks designed to operate in real-time, but these models still face challenges related to latency, retraining, and energy consumption. Unlike these approaches, our work emphasizes a balance between performance and feasibility by using protocol-level features combined with interpretable and lightweight models. This choice makes the proposed framework more suitable for integration into practical IoT scenarios where transparency, speed, and adaptability are critical.

Overall, these studies underline the growing role of ML techniques in addressing challenges such as energy optimization, routing efficiency, and network security in IoT systems. Building on these advancements, our work focuses on leveraging hybrid ML approaches to balance performance across multiple metrics, particularly in RPL-based IoT networks.

III. PROPOSED METHOD

The suggested approach uses machine learning methods to create an enhanced intrusion detection system (IDS) by exploiting the extensive dataset. Data preparation, feature selection, model training, and evaluation are some of the crucial processes in the methodology. The proposed framework consists of a full machine learning pipeline tailored to intrusion detection in IoT networks. It incorporates three classifiers for model training, Random Forest significance for feature selection, and preprocessing steps (imputation, normalization, and encoding of missing values). Flexible assessment of classification accuracy, robustness, and interpretability is made possible by this modular structure. Although supervised learning is used in this study, the framework can be extended to real-time or semisupervised architectures.

A. Data Preprocessing

The dataset first undergoes preprocessing in order to handle missing values, normalize numerical characteristics, and encode categorical variables. By ensuring that the data is clean and suitable for ML models, this raises the IDS overall accuracy and efficacy.

Let $X = \{x_1, x_2, ..., x_n\}$ represent the dataset, where each sample $x_i \in \mathbb{R}^d$ corresponds to a feature vector. Techniques like K-Nearest Neighbors are used to impute missing-data, while one-hot encoding is used to encode categorical features.

The normalization of numerical features is given by:

$$\hat{x}_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}, \quad j = 1, 2, \ldots, d$$

where μ_j and σ_j are the mean and standard deviation of feature x_j, respectively.

The datasets missing values were addressed using K-Nearest Neighbor imputation. This approach was selected because it estimates missing items using the most similar observations, preserving the local structure of the data. This enhances the quality of feature distributions for classification by maintaining significant links between attack patterns and protocol attributes in the context of network traffic data. To find the most pertinent factors for feature selection, Random Forest feautureimportance analysis was employed. This allowed us to reduce the feature space while retaining high-impact indicators such as tcprtt, dttl, and proto_tcp. This step not only enhanced model interpretability but also improved computational efficiency during training and evaluation.

B. Feature Selection

In order to decrease dimensionality and enhance model performance, feature selection is an essential stage in which the most pertinent features are found. Features that significantly aid in differentiating between malicious and legitimate traffic are chosen using methods like feature importance ratings and correlation analysis.

The feature importance Importance(X_j) of a feature X_j in a Random Forest model can be computed as:

$$\text{Importance}(X_j) = \sum_{t \in T} \Delta\text{Impurity}_t(X_j)$$

where T is the set of trees in the Random Forest, and $\Delta\text{Impurity}_t(X_j)$ is the reduction in impurity for feature X_j in tree t.

C. Model Training

To find the best model for intrusion detection, a variety of machine learning approaches are investigated. These include more sophisticated approaches like gradient boosting and deep learning models, as well as more conventional ones like decision trees, random forests, and support vector machines (SVM).

The goal of training a machine learning model is to learn the mapping $f : \mathbb{R}^d \to \{A, D, E, N\}$, where A, D, E, and N represent the attack categories (Analysis, DoS, Exploits, and Normal), and the model aims to minimize the following loss function:

$$L(\theta) = -\sum_{i=1}^{n} \sum_{c \in \{A,D,E,N\}} \mathbb{I}(y_i = c) \log P(y_i = c | x_i, \theta)$$

where: - $\mathbb{I}(y_i = c)$ is the indicator function, - $P(y_i = c|x_i, \theta)$ is the predicted probability that x_i belongs to class c.

D. Support Vector Machine (SVM) Training

The SVM model is formulated as follows for binary classification:

$$\min_{\mathbf{w},b,\xi} \left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \right)$$

subject to the constraints:

$$y_i(\mathbf{w} \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

where C is a regularization parameter, $\mathbf{w}$ is the weight vector, and ξ_i is the slack variable allowing for misclassification.

E. Cross-Validation

To ensure the reliability and generalization of the models, cross-validation methods are applied. The dataset is divided into k subsets. The model is then trained on $k-1$ subsets and tested on the remaining one. This procedure is repeated for each subset, and the overall performance is calculated as the average of the results obtained across all iterations.

IV. Security in IoT-Integrated Wireless Sensor Networks

In the context of the Internet of Things (IoT) [5], and more specifically within WSN [7], to stop unwanted access and lessen destructive activity, security must be mainained and intrusions must be detected. Packet-level metrics like destination TTL (dttl), source window size (swin), destination window size (dwin), TCP round-trip time (tcprtt), SYN-ACK packets (synack), and achnowledgment data packets (ackdat) are ampng the many features that the suggested system uses to detect possible threats. It also incorporates protocol-level information, such as the TCP (proto_tcp) and UDP (proto_udp) protocols, as well as servicespecific metrics like DNS (service_dns) requests. Furthermore, the system evaluates connection states, including active (state_CON) and closed (state_FIN) connections. Attack cat Analysis, attack cat Dos, attack cat Exploits, and attack cat Normal are the four primary categories into which the system divides intrusions based on the examinsation of these features: Analysis, Dos, Exploits, and Normal traffic. This multifaceted strategy improves the systems ability to identify breaches and preserve the security and integrity of IoTenabled WSNs, thereby protecting them from a wide range of cyberthreats. [15]

A. Attacks used in dataset

The assaults found in the dataset fail into four different categories: Normal, Exploits, Dos, and Analysis. A particular kind of harmful conduct directed towards computer systems is represented by each category. Belwo is a synopsis of every category [12]–[15]:

- **Analysis:** Activities intended to analyze a computer systems architecture and weaknesses fall under this category of assaults. Attackers may utilize this data to obtain sensitive intelligence or to develop more complex operations.
- **DoS (Denial of Service):** Dos attacks aim to deplete a computer systems or networks resources in order to prevent authorized users from accessing itUsing techniques like increasing network bandwidth, depleting system resources, or taking advantage of software flaws might cause this interruption.
- **Exploits:** Exploit attacks gain unauthorized access to a computer system by taking advantage of known or concealed flaws in operating systems or applications. These attacks give attackers the ability to execute malicious code,retrieve private data, or compromise the systems integrity.
- **Normal:** Network traffic that is benign and genuine and does not fall under any of the established attackcategories is included in this category. It acts as a starting point for distinguishing between malicious and legitimate traffic.

The following table presents a statistical summary of the attacks listed in the dataset:

TABLE I
DISTRIBUTION OF ATTACK CATEGORIES AND THEIR OCCURRENCES IN THE DATASET.

Attack Category	Number of Occurrences
Analysis	500
DoS	1000
Exploits	750
Normal	2000

B. Study Analysis and Evaluation of Anomaly Detection Techniques in WSN

In the field of cybersecurity and network traffic analysis [16], such a data collection [11] is of crucial importance. By analyzing these variables, researchers and analysts can uncover patterns, trends, and anomalies in network traffic [5]. For example, determining which protocols are most frequently utilized might provide insight on the kinds of services and applications that are most widely used on the network. Analyzing source and destination ports and source and destination IP adresses can also show communication patterns between various networks entities. Variables mlike ‘tcp_flags‘ ca, indicate different transmission control indications used in TCP communications, while measures like ‘frame_len‘ and ‘udp_len‘ can reveal information about the amount of data packets that are transmitted [2].

By combining this information with classification labels available in the ”label” and ”tipo_ataque” columns [2], it becomes possible to build more robust intrusion detection models and threat prevention systems. Indeed, by training machine learning algorithms [13], [14], [18] on this data, it is conceivable to develop systems capable of automatically identifying and flagging suspicious or malicious activities on the network [25], thereby enhancing the overall security of IT infrastructures. To sum up, this dataset is a great tool for network traffic analysis and cybersecurity [26] research and development [9], opening the door for new developments in defense against online attacks.

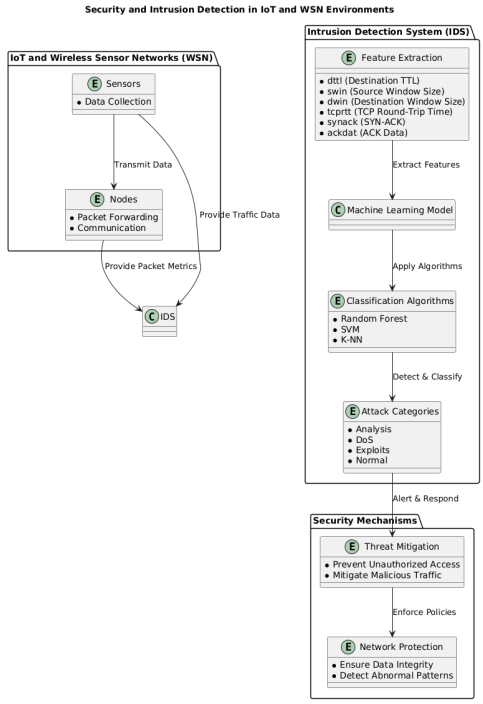


Fig. 1: Security and Intrusion Detection in IoT and WSN Environments

V. DATASET

With the rapid growth of the Internet of Things (IoT) and Wireless Sensor Networks, new forms of cyber threats are emerging that cannot be mitigated using conventional network security solutions. Machine learning (ML)-driven Intrusion Detection Systems (IDS) offer promising capabilities to address these challenges by detecting complex patterns of malicious behavior. In this study, we leverage a publicly available dataset derived from UNSW-NB15 [22], [23] and hosted on GitHub, specifically structured to reflect contemporary IoT traffic and attack scenarios. This dataset includes protocollevel and service-specific attributes that are rarely present in older benchmarks such as KDD99 or NSL-KDD. Our primary objective is to conduct a detailed comparative analysis of three well-known ML algorithms — Random Forest, SVM, and KNN — within a multi-class classification framework adapted to modern IoT network conditions. The proposed framework is designed to guide the development of scalable and reliable IDS solutions, particularly in smart city infrastructure and industrial IoT applications where detecting real-time threats is essential to maintaining service availability and data protection. The dataset used in this study was downloaded from a public GitHub repository [23] and was originally derived from UNSW-NB15. It includes pre-labeled instances representing different categories of network traffic (e.g., DoS, Exploits, Normal, Analysis). These labels were annotated in the original benchmark using a combination of automated logging and expert validation, ensuring class consistency across the dataset. This predefined labeling allows for multi-class classification experiments without the need for manual annotation during this study.

A. Description of Dataset Parameters

The dataset under study contains the following parameters:

- **Unnamed: 0** : Index of the row in the dataset, often generated automatically during data import.
- **dttl** : Time-to-Live (TTL) of the packets, representing the lifespan or number of hops a packet can make before being discarded.
- **swin** : Send window size in the TCP protocol, indicating the amount of data the sender is willing to send before receiving an acknowledgment.
- **dwin** : Receive window size in the TCP protocol, representing the amount of data the receiver is willing to accept.
- **tcprtt** : Round-Trip Time of TCP packets, measuring the delay between sending a packet and receiving the acknowledgment.
- **synack** : Time between sending a SYN (synchronize) packet and receiving the SYN-ACK (synchronize-acknowledge) packet in establishing a TCP connection.
- **ackdat** : Time between receiving an acknowledgment (ACK) and sending the corresponding data.
- **label** : Label indicating the nature of the traffic, such as normal or attacked.
- **proto_tcp** : Binary indicator (0 or 1) specifying if the TCP protocol is used.
- **proto_udp** : Binary indicator (0 or 1) specifying if the UDP protocol is used.
- **service_dns** : Binary indicator (0 or 1) specifying if the DNS service is used.
- **state_CON** : Binary indicator (0 or 1) specifying if the connection state is ”established” (CON).
- **state_FIN** : Binary indicator (0 or 1) specifying if the connection state is ”finished” (FIN).
- **attack_cat_Analysis** : Binary indicator (0 or 1) specifying if the attack belongs to the ”Analysis” category.
- **attack_cat_DoS** : Binary indicator (0 or 1) specifying if the attack belongs to the ”DoS” (Denial of Service) category.
- **attack_cat_Exploits** : Binary indicator (0 or 1) specifying if the attack belongs to the ”Exploits” category.
- **attack_cat_Normal** : Binary indicator (0 or 1) specifying if the traffic is normal.

The dataset utilized in this study consists of exactly **81,173 labeled instances** and **17 features**, each capturing critical protocol-level and service-specific attributes, such as ‘tcprtt‘, ‘dttl‘, and ‘proto_tcp‘. Three characteristics offer comprehensive details about every data sample, facilitating in-depth network traffic analysis and the identification of different

types of assaults. Specificaly, the dataset is categorized into four classes: *Normal, DoS, Exploits*, and *Analysis*. These labels were pre-assigned based on traffic fingerprints and packet behaviors identified during controlled simultaions scenarios, as documented in the source repository. Prior to public release, all IP addresses and protocol identifiers were anonymized using standard data masking techniques to ensure privacy and avoid leakage of sensitive information. This approach preserves the structural integrity of the data while supporting ethical handling of network traces for cybersecurity research.

VI. Simulation Results

A. Experiment environment

The tests were carried out on a device equipped with a CPU featuring the following specifications: an Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.70 GHz processor, 8.00 GB of RAM, and operating on a 64-bit Windows system. The Python programming language was utilized to create several categorization methods on Jupyter Notebook. Many Python libraries, including pandas(1.5.3), matplotlib, numpy, seaborn were used. Anaconda is responsible for installing these dependencies and tools.

B. SVM Results

The confusion matrix (Fig. 2) and ROC curve (Fig. 3) for the SVM model demonstrate its effectiveness in classifying multiple classes in the dataset. The confusion matrix highlights strong classification performance, with most predictions accurately aligned along the diagonal, such as 7759 for Class 5 and 3864 for Class 6. However, some misclassifications are observed, particularly in Classes 4 and 8. The ROC curve further validates this performance, showing excellent AUC values for most classes, including 1.00 for Classes 0, 2, 3, and 6, while Class 8 exhibits the lowest AUC of 0.49, indicating challenges in distinguishing this class. Overall, the SVM model performs exceptionally well for most categories, with room for improvement in specific cases.

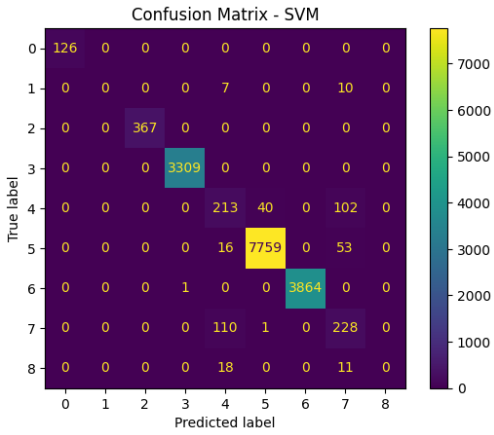


Fig. 2: Confusion Matrix of SVM

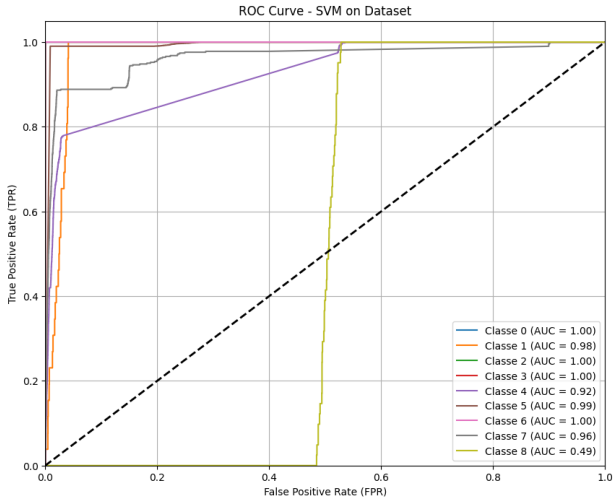


Fig. 3: ROC Curves of SVM

D. K-NN Results

The Random Forest model demonstrates strong overall performance in multiclass classification, as shown by the ROC curve (Fig. 5) and the confusion matrix. The ROC curve highlights excellent discrimination for most classes, with perfect AUC scores of 1.00 for classes 0, 2, 3, 5, and 6. However, moderate performance is observed for some classes, such as class 1 (AUC = 0.78) and class 8 (AUC = 0.66), indicating potential areas for improvement. The confusion matrix (Fig. 4) further confirms the model's effectiveness, with a high number of correct predictions along the diagonal, such as 3309 for class 3 and 7769 for class 5. Nonetheless, some misclassifications are evident, particularly between classes 4 and 7. These results suggest that while the model achieves excellent overall performance, further refinement is needed to address errors in more challenging or less represented classes.

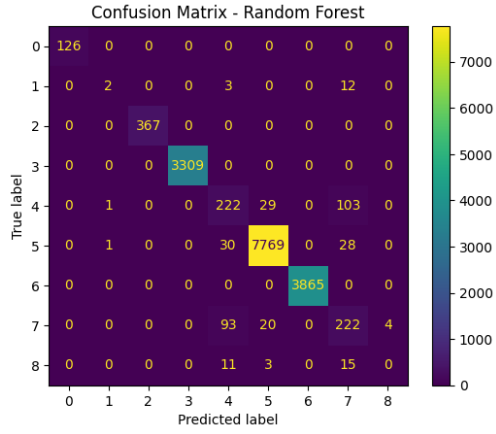


Fig. 4: Confusion Matrix of Random Forest

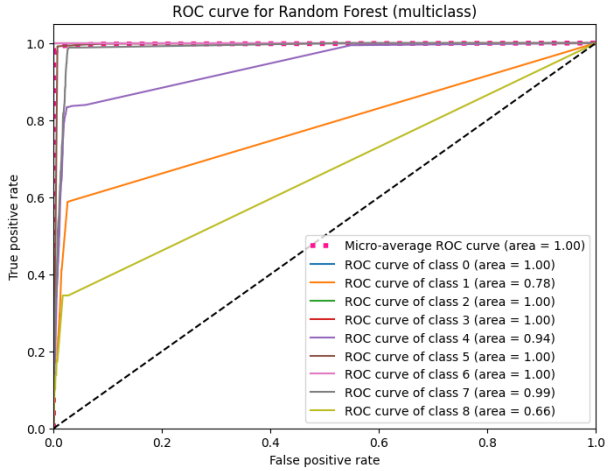


Fig. 5: ROC Curves of Random Fotrest

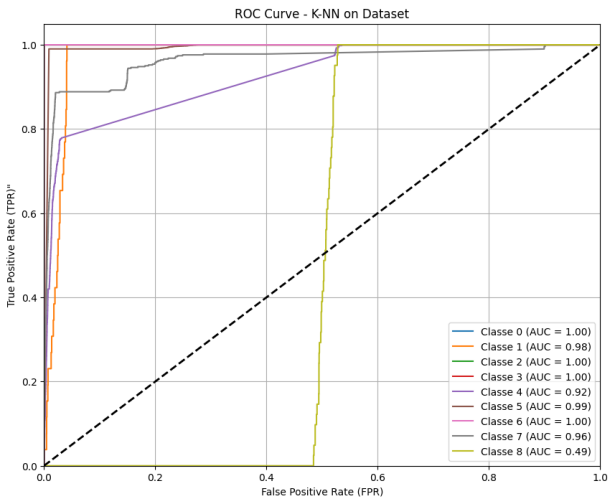


Fig. 7: ROC Curves of K-NN

D. K-NN Results

The confusion matrix (Fig. 6) demonstrates the K-NN algorithm's strong classification performance, with most predictions correctly aligned along the diagonal, such as 7780 for Class 5 and 3863 for Class 6. However, minor misclassifications are observed, like some instances of Class 3 being predicted as Class 4 or Class 7. These results indicate high overall accuracy, with slight areas for improvement in distinguishing closely related classes. The ROC curve (Fig. 7) for the K-NN algorithm illustrates the performance across multiple classes in terms of their True Positive Rate (TPR) and False Positive Rate (FPR). Most classes, such as Classes 0, 2, 3, and 6, achieve an AUC of 1.00, indicating perfect classification capability. Classes 5 and 7 also perform well with AUC values of 0.99 and 0.96, respectively, reflecting high accuracy. However, Class 8 stands out with an AUC of 0.49, suggesting significant challenges in distinguishing this class. Overall, the K-NN model demonstrates excellent classification performance for most classes, but there is room for improvement in handling specific cases like Class 8.

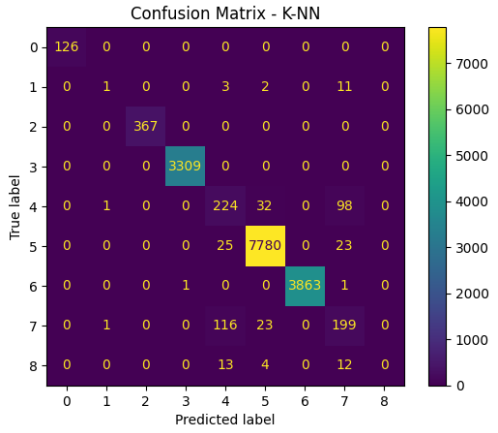


Fig. 6: Confusion Matrix of K-NN

E. Performance Metrics Definitions

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Where:

- TP stands for True Positives,
- TN stands for True Negatives,
- FP stands for False Positives,
- FN stands for False Negatives.

Similarly, precision and F1 score can be defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Where:

- Precision measures the proportion of true positive predictions out of all positive predictions made.
- Recall measures the proportion of true positive predictions out of all actual positive instances in the data.

F. Comparison of Performance Metrics Across Algorithms

The bar plots in Fig. 8, Fig. 9, Fig. 10 compare the performance of three machine learning algorithms—K-NN, Random Forest, and SVM—using the F1-score, precision, and recall metrics. For all metrics, the weighted averages outperform macro averages, reflecting the algorithms' ability to perform well across all classes, even in imbalanced datasets. Among the algorithms, Random Forest consistently achieves the highest scores across all metrics, indicating its robustness and reliability for classification tasks. SVM and K-NN also demonstrate strong performance, with slight variations in macro and weighted

averages. These results highlight Random Forest's superior capability in maintaining a balance between precision, recall, and F1-score, making it the most effective model for the dataset.

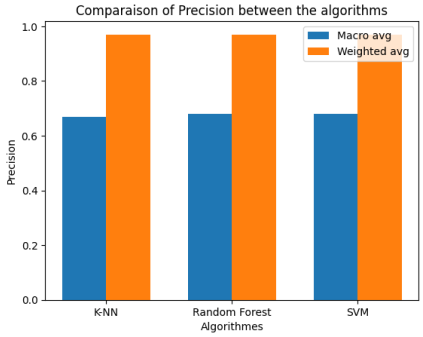


Fig. 8: Comparison with precision

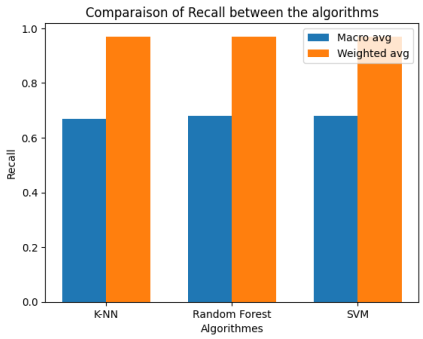


Fig. 9: Comparison with recall

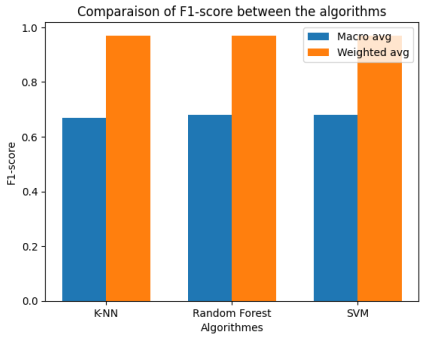


Fig. 10: Comparison with f1-score

G. *Learning Curve Analysis for SVM, K-NN, and Random Forest Models*

The learning curves for the SVM (Fig. 11), K-NN (Fig. 12), and Random Forest (Fig. 13) models provide valuable insights into their performance during training and testing as the dataset size increases.

H. *SVM Learning Curves*

- The training accuracy gradually improves and stabilizes around 97.6%, demonstrating that the SVM model effectively fits the training data.
- The testing accuracy closely follows the training accuracy, with a slight gap indicating good generalization and minimal overfitting.
- The consistent improvement in both curves as more data is added highlights the SVM model's ability to generalize well across different dataset sizes.

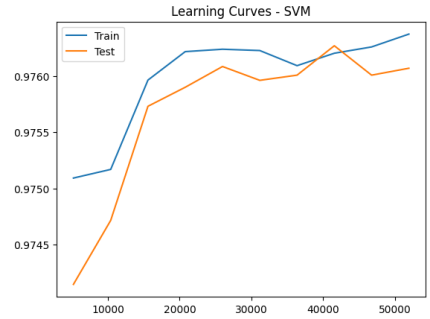


Fig. 11: SVM Learning Curves

I. *K-NN Learning Curves*

- The training accuracy starts high and stabilizes at approximately 98.4%, indicating effective fitting to the training data.
- The testing accuracy improves gradually, stabilizing around 97.6%. However, the slight gap between training and testing accuracies suggests mild overfitting.
- K-NN benefits significantly from larger datasets, as shown by the steady improvement in testing accuracy with increased data size.

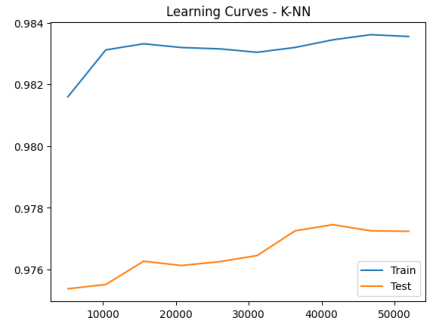


Fig. 12: k-NN Learning Curves

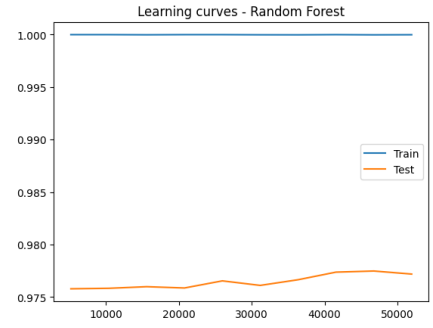


Fig. 13: RF Learning Curves

J. *Random Forest Learning Curves*

- The training accuracy remains at a perfect 100% across all dataset sizes, reflecting the Random Forest model's high capacity to fit the training data.

- The testing accuracy stabilizes around 98.9%, with a minimal gap from the training curve, indicating excellent generalization and low overfitting.
- Random Forest demonstrates robustness and high reliability, achieving strong performance even with smaller datasets.

K. Key Insights

- **Random Forest** outperforms both SVM and K-NN in terms of generalization, as evidenced by the minimal gap between training and testing accuracies.
- **SVM** exhibits strong generalization capabilities, with consistent performance improvements as the dataset size increases.
- **K-NN** shows slight overfitting but remains effective, benefiting from larger datasets for improved generalization.
- These results highlight Random Forest as the most robust model, followed by SVM, while K-NN requires more data to bridge the training-testing accuracy gap.

L. Performance Evaluation of Machine Learning Algorithms

Three machine learning [27] algorithms: Random Forest, KNN and SVM are evaluated, and the results demonstrate their remarkable dependability and performance in the task at hand. With accuracy, precision, recall, and F1-score all at 99.9877%, Random Forest produced impressive results, proving its dependability and low categorization errors.With all metrics, accuracy, precision, recall, and f1 score at 99.9754%, KNN also demonstrated remarkable performance. Its robustness and accuracy in classifying cases are highlighted by this constancy. With an accuracy of 99.9630% and precision, recall, and F1 scores that closely matched at 99.9630%, SVM produced impressive reesults although somewhat lagging behind the other two. SVM is a reliable option for applications needing high precision and consistency because of three results, which validate its efficacy and balance.

TABLE II
PERFORMANCE METRICS OF CLASSIFICATION MODELS.

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.999877	0.999877	0.999877	0.999877
K-Nearest Neighbors	0.999754	0.999754	0.999754	0.999754
SVM	0.999630	0.999631	0.999630	0.999630

VII. STATISTICAL SIGNIFICANCE ANALYSIS OF CLASSIFIER PERFORMANCE

To statistically validate the performance differences among the classifiers, we conducted the Wilcoxon signed-rank test on the 5-fold cross-validation results. As shown in Table III, the p-values for all pairwise comparisons between Random Forest, SVM, and KNN exceeded the 0.05 threshold, indicating that none of the observed differences were statistically significant. This suggests that the variations in accuracy are consistent across folds and not due to random chance. While Random Forest achieved slightly higher average scores, the results confirm that all models perform competitively on this dataset, reflecting its balanced structure and the robustness of the feature engineering pipeline.

TABLE III
WILCOXON SIGNED-RANK TEST FOR CLASSIFIER PERFORMANCE COMPARISON

Algorithms	p-value	Significance (p ¡ 0.05)
Random Forest vs SVM	0.1250	No
Random Forest vs KNN	0.0625	No
SVM vs KNN	0.1250	No

VIII. DISCUSSION OF THE RESULTS

This study examines the performance of three machine learning algorithms—Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—on a specific classification task. The results highlight the outstanding effectiveness of these algorithms, as evidenced by their high-performance metrics, confirming their reliability and suitability for the task. The Random Forest algorithm achieved an accuracy of 99.9877%, with precision, recall, and F1 score all matching at 99.9877%. This indicates an exceptional level of reliability, with the algorithm showing minimal classification errors. The consistent high scores across all metrics reflect the algorithm's capacity to accurately identify both positive and negative cases, making it particularly effective for scenarios where precise classification is critical.

Similarly, the K-Nearest Neighbors algorithm exhibited strong performance, with an accuracy, precision, recall, and F1 score of 99.9754%. Although slightly lower than Random Forest, these metrics still highlight KNN's robustness as a classifier. The consistency in its performance metrics suggests that KNN is highly reliable and capable of accurately classifying instances, making it a viable alternative in cases where simplicity and interpretability are favored.

The Support Vector Machine algorithm, while performing marginally below the other two, still demonstrated impressive results. With an accuracy of 99.9630%, precision of 99.9631%, recall of 99.9630%, and F1 score of 99.9630%, SVM shows a strong balance between precision and recall. These metrics highlight its effectiveness and reliability, suggesting that SVM is well-suited for tasks that demand high accuracy and consistent performance. While the reported performance metrics are extremely high, care was taken to mitigate overfitting. Additionally, while the dataset used in this study is relatively balanced in terms of labeled categories, some subtle internal imbalances could still influence model sensitivity to rare attack types. Further work could explore resampling techniques or class-weighted training to better address these effects in critical classification scenarios. Learning curves were analyzed for each classifier, and the gaps between training and validation accuracy remained minimal, suggesting that the models generalized well to unseen data. Additionally, the use of 5-fold cross-validation helped ensure the robustness of the results. Future versions of this work may include statistical tests such as paired t-tests or Wilcoxon signed-rank tests to assess

the significance of observed differences between classifiers, especially when performance metrics are very close. While the models demonstrated strong performance on the offline dataset, we acknowledge that real-world deployments involve continuous and dynamic data streams. Our current evaluation is limited to static scenarios, which may not fully capture temporal drift or evolving attack behaviors. Furthermore, the dataset, although balanced at a class level, may still contain subtle distributional imbalances that affect the learning process. We plan to extend this work by incorporating realtime data handling and deeper analysis of class distribution effects, particularly in detecting rare or stealthy attack patterns. Although our study relies on well-known machine learning algorithms such as Random Forest, SVM, and KNN, the novelty lies in the integration of these models with protocol-level and service-specific attributes rarely leveraged in existing IDS literature. By using a structured multi-class dataset tailored for IoT intrusion scenarios, we offer a new perspective that goes beyond traditional flow-based detection. This combination enhances both interpretability and detection precision in IoT environments, particularly in edge-constrained WSNs.

In conclusion, the Random Forest method performs exceptionally well overall, however it differs slightly from the other two algorithms in terms of accuracy and consistency. Nonetheless, depending on the particular needs of the assignment, KNN and SVM can also offer strong and reliable classification, making them appropriate alternatives. These algorithms' excellent performance highlights their potential for practical implementation in real-world categorization issues, laying a solid basis for additional study and application.

IX. CONCLUSION

In conclusion, this study proposes a comprehensive approach to developing an advanced intrusion detection system (IDS) leveraging machine learning algorithms. By employing systematic data preprocessing, strategic feature selection, and rigorous model evaluation, the methodology demonstrates remarkable effectiveness in detecting and classifying network intrusions. The preprocessing phase ensures the dataset is prepared for model training by addressing issues such as missing values, normalization, and encoding, laying a strong foundation for accuracy and efficiency. Feature selection further optimizes the dataset by identifying the most relevant attributes, reducing dimensionality, and enhancing model performance. Multiple machine learning algorithms, including Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), were evaluated for their intrusion detection capabilities. The results highlight their high reliability and robustness, with Random Forest achieving an exceptional accuracy of 99.9877%. KNN and SVM also delivered strong performances, with accuracies of 99.9754% and 99.9630%, respectively, demonstrating their potential as reliable alternatives based on specific needs. The thorough evaluation using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, combined with cross-validation techniques, ensures the robustness and generalizability of the models. This comprehensive assessment not only highlights the strengths and weaknesses of each model but also provides valuable insights into their capabilities in detecting various attack categories. While the reported performance metrics are extremely high, care was taken to mitigate overfitting. Learning curves were analyzed for each classifier, and the gaps between training and validation accuracy remained minimal, suggesting that the models generalized well to unseen data. Additionally, the use of 5-fold cross-validation helped ensure the robustness of the results. Moreover, future iterations of this framework could incorporate hybrid models that combine decision-tree-based learners with deep learning layers to capture both hierarchical structure and temporal dependencies, thus enhancing detection robustness in evolving IoT environments. Future versions of this work may include statistical tests such as paired ttests or Wilcoxon signed-rank tests to assess the significance of observed differences between classifiers, especially when performance metrics are very close. This work highlights the overlooked value of protocol-level features for achieving interpretable and precise intrusion detection. These results confirm that protocol-level features enable efficient and interpretable IDS. Future work will explore hybrid models to enhance realtime adaptability. Overall, the proposed method enhances the detection capabilities of IDS by leveraging advanced machine learning techniques. The high performance of the evaluated algorithms indicates their potential for practical deployment in real-world network security applications. This study lays a solid foundation for future research and development in the field of intrusion detection, contributing to the broader goal of improving cybersecurity measures and protecting against evolving cyber threats.

REFERENCES

[1] Salim El Khediri, Awatef Benfradj, Adel Thaljaoui, Tarek Moulahi, Rehan Ullah Khan, Abdullatif Alabdulatif, and Pascal Lorenz, "Integration of artificial intelligence (AI) with sensor networks: Trends, challenges, and future directions," *Journal of King Saud University - Computer and Information Sciences*, 2024. **DOI**: 10.1016/j.jksuci.2023.101892.

[2] Gutierrez-Portela Fernando, Arteaga-Arteaga Harold Brayan, Almenares Mendoza Florina, Calderón-Benavides Liliana, Acosta-Mesa Héctor-Gabriel, and Tabares-Soto Reinel, "Enhancing intrusion detection in IoT communications through ML model generalization with a new dataset (IDSAI)," *IEEE Access*, 2023. **DOI**: 10.1109/ACCESS.2023.3292267.

[3] C. N. Vanitha, S. Malathy, Rajesh Kumar Dhanaraj, and Anand Nayyar, "Optimized pollard route deviation and route selection using Bayesian machine learning techniques in wireless sensor networks," *Computer Networks*, 2023. **DOI**: 10.1016/j.comnet.2022.109228.

[4] Dharini N, Jeevaa Katiravan, Sruthi Priya D. M., and Sakthi Sneghaa V. A., "Intrusion detection in novel WSN-Leach DoS attack dataset using machine learning-based boosting algorithms," *Procedia Computer Science*, 2023. **DOI**: 10.1016/j.procs.2023.12.064.

[5] Marouane Myyara, Oussama Lagnfdi, Anouar Darif, and Abderrazak Farchane, "Ehancing QoS for IoT Devices through Heuristics-based Computation Offloading in Multi-access Edge Computing," *InfoCommunications*, 2024. **DOI**: 10.36244/ICJ.2024.4.2.

[6] S. Lakshmi Narayanan, M. Kasiselvanathan, K.B. Gurumoorthy, and V. Kiruthika, "Particle swarm optimization-based artificial neural network (PSO-ANN) model for effective k-barrier count intrusion detection system in WSN," *Measurement: Sensors*, 2023. **DOI**: 10.1016/j.measen.2023.100875.

[7] Iqbal Jebril, M. Premkumar, Ghaida Muttashar Abdulsahib, S. R. Ashokkumar, S. Dhanasekaran, Oshamah Ibrahim Khalaf, and Sameer Algburi, "Deep Learning based DDoS Attack Detection in Internet of Things: An Optimized CNN-BiLSTM Architecture with Transfer Learning and Regularization Techniques," *InfoCommunications*, 2024. **DOI**: 10.36244/ICJ.2024.1.1.

[8] I. Surenther, K. P. Sridhar, and Michaelraj Kingston Roberts, "Enhancing data transmission efficiency in wireless sensor networks through machine learning-enabled energy optimization: A grouping model approach," *Alexandria Engineering Journal*, 2024. **DOI**: 10.1016/j.asej.2024.102644.

[9] Saziya Tabbassumand Rajesh Kumar Pathak, "Effective data transmission through energy-efficient clustering and fuzzy-based IDS routing approach in WSNs," *Visual Informatics*, 2024. **DOI**: 10.1016/j.vrih.2022.10.002.

[10] I. Surenther, K. P. Sridhar, and Michaelraj Kingston Roberts, "Enhancing data transmission efficiency in wireless sensor networks through machine learning-enabled energy optimization: A grouping model approach," *Alexandria Engineering Journal*, 2024. **DOI**: 10.1016/j.asej.2024.102644.

[11] Kai-Yun Tsao, Thomas Girdler, and Vassilios G. Vassilakis, "A survey of cyber security threats and solutions for UAV communications and flying ad-hoc networks," *Ad Hoc Networks*, 2022. **DOI**: 10.1016/j.adhoc.2022.102894.

[12] Trinh Thuc Lai, Tuan Phong Tran, Jaehyuk Cho, and Myungsik Yoo, "DoS attack detection using online learning techniques in wireless sensor networks," *Alexandria Engineering Journal*, 2023. **DOI**: 10.1016/j.aej.2023.11.022.

[13] Ayuba John, Ismail Fauzi Bin Isnin, Syed Hamid Hussain Madni, and Muhammed Faheem, "Cluster-based wireless sensor network framework for denial-of-service attack detection based on variable selection ensemble machine learning algorithms," *Intelligent Systems with Applications*, 2024. **DOI**: 10.1016/j.iswa.2024.200381.

[14] Bhanu Chander and G. Kumaravelan, "Outlier detection strategies for WSNs: A survey," *Journal of King Saud University - Computer and Information Sciences*, 2022. **DOI**: 10.1016/j.jksuci.2021.02.012.

[15] Suriyan Kannadhasan and Ramalingam Nagarajan, "Intrusion detection in machine learning-based E-shaped structure with algorithms, strategies, and applications in wireless sensor networks," *Heliyon*, 2024. **DOI**: 10.1016/j.heliyon.2024.e30675.

[16] S. Anitha, S. Saravanan, and A. Chandrasekar, "Trust management-based multidimensional secure cluster with RSA cryptography algorithm in WSN for secure data transmission," *Measurement: Sensors*, 2023. **DOI**: 10.1016/j.measen.2023.100889.

[17] J. Cevallos, A. Shaghaghi, and S. Nepal, "Few-shot Zero-Day Attack Detection for IoT Networks," in *Proc. IEEE Intl. Conf. on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2023, pp. 371–378. **DOI**: 10.1109/TrustCom57881.2023.00064.

[18] M. Ejaz, M. M. Rathore, A. Paul, and S. W. Kim, "Real-Time Intrusion Detection in IoT Networks Using Edge Computing," *Sensors*, vol. 19, no. 10, pp. 2356–2370, May 2019. **DOI**: 10.3390/s19102356.

[19] A. Roy, A. Ghosh, and R. Buyya, "Federated Learning for Privacy-Preserving Intrusion Detection in Industrial IoT: The 2DF-IDS Framework," *ACM Trans. Internet Technol. (TOIT)*, vol. 23, no. 1, pp. 1–25, Jan. 2023. **DOI**: 10.1145/3571862.

[20] Y. Liu, S. Leng, X. Zhang, and K. Yang, "Context-Aware Intrusion Detection in Internet of Vehicles Based on Multi-Agent Reinforcement Learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 756–769, Jan. 2022. **DOI**: 10.1109/TVT.2021.3129956.

[21] N. Kumar and R. Ghosh, "Challenges and Research Opportunities for Intrusion Detection in IoT: A Comprehensive Review," *Comput. Commun.*, vol. 202, pp. 68–89, Dec. 2022. **DOI**: 10.1016/j.comcom.2022.07.017.

[22] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 Network Data Set)," in *Proc. Military Communications and Information Systems Conference (MilCIS)*, Canberra, Australia, Nov. 2015, pp. 1–6. **DOI**: 10.1109/MilCIS.2015.7348942.

[23] A. Bhardwaj, "IoT Network Intrusion Detection System – UNSW-NB15 Dataset," *GitHub repository*, 2021. [Online]. Available: https://github.com/abhinav-bhardwaj/IoT-Network-Intrusion-Detection-System-UNSW-NB15

[24] I. Jebril, M. Premkumar, G. M. Abdulsahib, S. R. Ashokkumar, S. Dhanasekaran, O. I. Khalaf, and S. Algburi, "Deep learning based DDoS attack detection in Internet of Things: An optimized CNN-BiLSTM architecture with transfer learning and regularization techniques," *Infocommunications Journal*, vol. XVI, no. 12, pp. 1–11, Mar. 2024. **DOI**: 10.36244/ICJ.2024.1.1.

[25] I. Larhlimi, M. Lmkaiti, M. Lachgar, H. Ouchitachen, and A. Darif, "Genetic Algorithm-Driven Cover Set Scheduling for Longevity in Wireless Sensor Networks," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 3, pp. 1104–1110, Mar. 2025. **DOI**: 10.14569/IJACSA.2025.01603135.

[26] M. Lmkaiti, M. Lachgar, I. Larhlimi, H. Moudni, and H. Mouncif, "Secure Optimization of RPL Routing in IoT Networks: Analysis of Metaheuristic Algorithms in the Face of Attacks," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 4, pp. 1184–1191, Apr. 2025. **DOI**: 10.14569/IJACSA.2025.01604138.

[27] M. Lmkaiti, H. Moudni, and H. Mouncif, "Machine learning–based detection in wireless sensor networks," *Recent Advances in Internet of Things Security*, pp. 19–29, 2025. **DOI**: 10.1201/9781003587552-3.

Mansour Lmkaiti is with the Department of Computer Mathematics, Polydisciplinary Faculty, University Sultan Moulay Slimane, Morocco. (ORCID ID: https://orcid.org/ 0009-0000-5882-3622). Domains of interest include highperformance computer systems and networks, machine learning algorithms, high performance in WSNs, and cybersecurity in wireless sensor networks.

Ibtissam Larhlimi obtained her B.Sc in Systems Informatiques Répartis from Université Qadi Ayyad, Marrakech, Morocco, in 2018, and her M.Sc in Telecommunications Systems and Computer Networks from Université Sultan Moulay Slimane, Beni Mellal, Morocco, in 2020. She is pursuing her PhD in mathematics and computer science at Sultan Moulay Slimane University, Beni Mellal, Morocco. Her research interests include maximizing the lifetime of RCSFs via cooperation between nodes in the context of the IoT.

Maryem Lachgar obtained her B.Sc in Systems Informatiques Répartis from Université Qadi Ayyad, Marrakech, Morocco, in 2018, and her M.Sc in Telecommunications Systems and Computer Networks from Université Sultan Moulay Slimane, Beni Mellal, Morocco, in 2020. She is pursuing her PhD in mathematics and computer science at Sultan Moulay Slimane University. Her research focuses on improving routing algorithms in wireless sensor networks based on Ultra Wideband.

Houda Moudni is currently an Assistant Professor at the National School of Business and Management, Sultan Moulay Slimane University, Beni Mellal, Morocco. She received her Ph.D. degree in Computer Sciences from the Faculty of Sciences and Technology of Beni Mellal in 2019. (ORCID ID: https://orcid.org/0009-0001-1746-9249.) Her research focuses on securing routing protocols in Mobile Ad Hoc Networks (MANET), Wireless Sensor Networks (WSN), and the Internet of Things (IoT).

Hicham Mouncif is a Full Professor and Ph.D. Supervisor in the Department of Computer Sciences, Polydisciplinary Faculty of Beni Mellal, University Sultan Moulay Slimane. He has published numerous academic papers in distinguished journals based on teaching and research experience. As the Coordinator of the Computer Systems Engineering Master's Program, his research interests include educational technologies, machine learning, transportation networking, and routing protocols.(ORCID ID: https://orcid.org/0000-0003-3312-8230)

Multi-photon QKD for Practical Quantum Networks

Nitin Jha, Abhishek Parakh, and Mahadevan Subramaniam

***Abstract*—Quantum key distribution (QKD) will most likely be an integral part of any practical quantum network in the future. However, not all QKD protocols can be used in today's networks because of the lack of single-photon emitters and noisy intermediate quantum hardware. Attenuated-photon transmission, typically used to simulate single-photon emitters, severely limits the achievable transmission distances and makes the integration of the QKD into existing classical networks, that use tens of thousands of photons per bit of transmission, difficult. Furthermore, it has been found that protocol performance varies with topology. In order to remove the reliance of QKD on single-photon emitters and increase transmission distances, it is worthwhile to explore QKD protocols that do not rely on single-photon transmissions for security, such as the 3-stage QKD protocol, which can tolerate multiple photons in each burst without information leakage. This paper compares and contrasts the 3-stage QKD protocol with conventional QKD protocols and its efficiency in different network topologies and conditions. Furthermore, we establish a mathematical relationship between achievable key rates to increase transmission distances in various topologies..**

***Index Terms*—Quantum networks, quantum key distribution, multi-photon transmissions, 3-stage protocol, network topologies.**

I. INTRODUCTION

Quantum information theory has been of significant interest over recent decades, leading to advances in computing, networking, and sensing technologies [1]–[3]. Quantum networks, under specific hardware configurations, offer enhanced security for communications [4] and are seen as a vital element in the future development of the quantum internet. Despite some challenges [5], Quantum Key Distribution (QKD) remains a foundational technique for ensuring network security. Research efforts are currently directed towards developing near-ideal quantum hardware, perfecting QKD systems, and mitigating potential eavesdropping risks [6]–[8]. Practical implementations have been demonstrated through the DARPA network and other networks across Europe [9] and Tokyo [10], showing the feasibility of quantum networks. Recent developments include the deployment of a private quantum network across Europe during the 2021 G20 Summit in Trieste, featuring two sender and two receiver nodes [11]. Furthermore, from 2003 to 2016, the Chinese Academy of Sciences (CAS) developed a satellite and ground-based network, achieving secure key distribution over $7,600$ km using a decoy-state QKD transmitter onboard a low-earth-orbit satellite, which also facilitated a secure video conferencing session [12]. These advancements underscore the rapid progress in quantum technologies that are poised to support large-scale secure quantum communications [8], [13].

QKD offers unconditional security because it utilizes the principles of quantum physics, making it *future proof* against eavesdropping, unlike classical transmission methods [14]–[18]. QKD, combined with *quantum-resistant classical algorithms* and *quantum cryptography*, is expected to play a crucial role in securing future communications for applications such as smart grids and defense networks [19]. Despite its potential, QKD faces several challenges, both theoretical and practical. These include photon number splitting (PNS) attacks, hardware inconsistencies [20], [21], and vulnerabilities related to Bell inequality test *loopholes* [22]. Furthermore, practical issues with optical switches and trusted nodes limit network robustness and range, with compromised nodes posing significant security risks [23]. Current approaches aim to manage multi-photon bursts, but these methods are limited because they still discard multi-photon bursts and rely on single photons, making integration with classical networks challenging. An alternative is the use of advanced multiphoton QKD protocols, such as the 3-stage protocol, which do not depend on single photons and accommodate higher photon burst rates [24]–[26]. Recent studies have also shown potential for using the three-stage protocol in quantum secure direct communication [27], [28]. This investigation forms the core of our study. Some other experimental works show the use of multiphoton approach for several quantum communication setups. Such as, [29] proposes a modified, more efficient twin-field (TF) QKD using a sending-or-not-sending (SNS) TF-QKD, and also studies the case of two-photon emissions instead of single photon QKD. Furthermore, [30] studies coincidence measurements to monitor and selectively include multi-photon pulses (two- and three-photon events) in the key generation—rather than discarding them as in standard BB84—it is possible to boost the secure key rate by roughly 74% over line-of-sight channels without resorting to decoy states. [31] studies the effect of different noise models on quantum memory of quantum repeaters which would be central to development of large scale quantum networks. [32] reviews the progress made in field of free-space QKD networks which highlights some of the key-experimental work done in deploying free-space QKD networks. Many of the fundamental works in field of QKD are theoretical and simulation based, while some provides experimental validation to these theoretical works. In this study, we simulate quantum networks based on several practical network parameters such as attenuation coefficient, noise parameters, etc.

Nitin Jha and Abhishek Parakh are with the Department of Computer Science Kennesaw State University, Kennesaw, Georgia, USA (e-mail: njha1@students.kennesaw.edu, aparakh@kennesaw.edu)

Mahadevan Subramaniam is with the Computer Science Department University of Nebraska Omaha, Nebraska, USA (e-mail: msubramaniam@unomaha.edu)

DOI: 10.36244/ICJ.2025.2.9

This paper looks at three different QKD protocols, i.e., the Decoy-state, the 3-stage, and the E91 protocols. Our study describes the efficiencies of the above protocols on different topologies such as direct, line, grid, ring, and torus topology. We vary network parameters such as entanglement swapping success probability, decoherence probability, and signal attenuation during transmission. We move on to analyze in detail the performance of E91 and the 3-stage protocol on the torus topology. To establish the significance of the multi-photon bursts in current practical scenarios, we analyze multi-photon bursts up to a burst size of a million qubits. The multiphoton burst represents generating a burst of a given size, and then encoding them as a quantum state. This allows for measurements, and thus collapsing the final result into on classical bit. This analysis led to defining a mathematical relationship between the size of the multi-photon burst used and the maximum distance of stable transmission between Alice and Bob. The preprint of this paper is located at [33].

II. QKD PROTOCOLS

QKD protocols can be classified based on the detection techniques used to retrieve the key information encoded in the photons being used. Discrete-Variable (DV) protocols use the polarization (or phase) of weak coherent pulses to encode the information, which simulates a true single-photon state [19]. Protocols such as Decoy-state and BB84 use the single photon-encoding scheme where the information is encoded in the polarization of the photon being used. Another category of protocols is called Continuous-Variable QKD, a technique where photon-counters are replaced with general p-i-n photo-diodes, which are known to be faster [34]. The detection techniques used in the above are based on *"homodyne detection"*.

The *Decoy-state* protocol enhances BB84 security by using random photon bursts to counteract PNS attacks, preventing Eve from distinguishing real from decoy transmissions [17]. Implemented using weak coherent pulses, this protocol varies the photon burst intensity and uses statistical analysis to detect intrusions [35], [36]. Post-transmission, Alice and Bob verify detection parameters over an authenticated channel to estimate and detect any attacker presence [37]. The *three-stage* protocol is a viable option for practical quantum networks and it uses a technique similar to classical *double-lock encryption* [24], [38]. In this protocol, Alice starts by encoding a key or message using orthogonal quantum states or an arbitrary state $|X\rangle$. Both Alice and Bob apply secret unitary operations, U_A and U_B, that commute ($[U_A, U_B] = 0$) [24], [38]. The three-stage protocol is shown to be multiphoton resilient. This works by encoding multiple qubits for each classical bit, and since the preparation basis is global knowledge, all qubits associated with one bit are prepared in the same global basis. This multiphoton implementation of the three-stage protocol has been shown to be a viable option for sending direct messages across the quantum channel. [39] Fig.(1) represents the working of the three-stage protocol introduced by [24]. The *E91 protocol*, unlike the BB84 and B92 protocols, is an entanglement-based protocol utilizing single-photon transmissions. In our implementation, the $E91$ protocol uses the entanglement distribution through quantum repeater nodes through the entanglement distribution. This can be interpreted as distributing multipartite entangled state across each pulse train. In post-processing, Alice and Bob keep only those coincidence events corresponding to one photon at each end; any higher-order multiphoton coincidences are treated as erasures. This makes E91 more efficient over longer distances, as quantum repeaters help in the reduction of the attenuation loss.

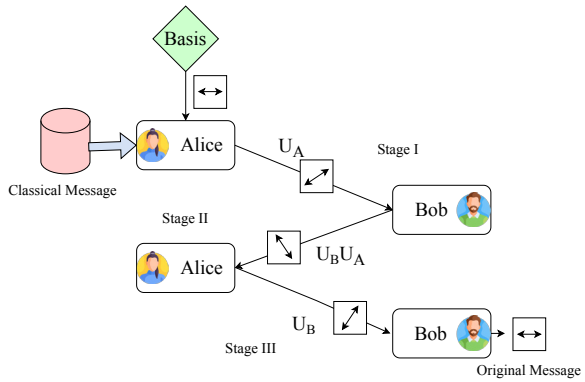


Fig. 1: Schematic diagram representing the working of the three-stage protocol.

The main security concern with multi-photon transmissions is the PNS attack. We can define the probability of finding $n-$coherent photons as the Poisson distribution [40],

$$P(n,\mu) = \frac{\mu^n e^{-\mu}}{n!}, \tag{1}$$

where μ is the mean photon number for the photon burst. The zero, first, and second-order expansion for our photon distribution are,

$$P(0) \approx 1 - \mu + \frac{\mu^2}{2};\ P(1) \approx \mu - \mu^2;\ P(2) \approx \frac{\mu^2}{2}. \tag{5}$$

Now, for a burst containing more than 1 photon, i.e., $n \geq 2$, we can write the Poisson distribution for photons as [40],

$$P(n \geq 2) \approx \frac{\mu}{2} + \frac{\mu^2}{4}. \tag{6}$$

The two main strategies employed by the eavesdropper to compromise the security of quantum communications are [40]:

- The first strategy involves Eve intercepting and analyzing all photons sent by Alice, and then relaying plausible states to Bob through a nearby source, carefully mimicking photon statistics. This scenario reveals that the maximum ratio of mutual information between Eve and Alice to the Quantum Bit Error Rate (QBER) reaches 6.83. Under *infinitesimal splitting*, it's shown that Eve could achieve complete informational equivalence with Alice, indicated by a mutual information score of 1.
- The second strategy involves Eve using a beamsplitter to extract a fraction λ from each pulse, simultaneously reducing photon loss by replacing the line with one of

lower loss. The mutual information between Alice and Eve in this scenario is quantified as,

$$I(A,E) = \frac{\mu}{2}(2\lambda(1-\lambda))\frac{1}{2}, \tag{7}$$

where $\mu/2$ represents the probability $P(2)$. This model predicts a maximum information gain for Eve of $\mu/8$ when $\lambda = 1/2$, correlating to a 3 dB gain.

The key rate and the distance over which these key rates are stable are highly correlated to the presence of an eavesdropper in the system. The efficiency of the key distribution is highly dependent on the Quantum Bit Error Rate (QBER) and the mutual information between Alice and Eve, $I(A,E)$. With today's equipment, the major problem is the higher detector noise, which leads to high QBER at large distances; thus, the maximum distance of stable transmission drastically decreases [40].

One important aspect of developing efficient quantum networks is the efficiency of transmissions in different network topologies. Apart from studying some of the basic topologies, such as *line*, *star*, *ring*, and *grid*, we designed a *torus topology* for our network simulator. Table I compares the advantages and disadvantages of the various topologies used in this study. Figure (2) shows the different topologies generated through our network simulator.

III. Network Simulation Setup

The Network Simulator developed for this study was written in Matlab. The network simulator was enhanced using the simulator presented in [41].

A. Network Communication

The connections between each of the involved nodes are achieved by simulating fiber-optic cables with transmission probability as given in equation(8).

$$P(L) = 10^{-\alpha L/10}, \tag{8}$$

where L is the fiber length in kilometers, and α is the attenuation coefficient of the fiber-optic cable used. It's evident from equation(8) that as we increase the distance between the nodes or the length of fiber required, the success of transmission would drastically decrease. For our simulation, we keep the value of $\alpha = 0.15\ km^{-1}$ unless otherwise specified. Our network simulation includes various types of nodes, with trusted nodes like Alice and Bob being crucial. These nodes attempt to connect through viable paths, storing successfully transmitted qubits in a raw key pool, $(RK)_{i,j}$, where i,j represents the node pair. The simulation also incorporates optical switches for linking nodes on a peer-to-peer basis, which, while facilitating dynamic connections, show performance deterioration. Further, we utilize the E91 Protocol to simulate quantum repeaters. Unlike classical repeaters, quantum repeaters use entanglement and entanglement swapping to extend the communication range without violating the no-cloning principle of quantum mechanics. The effectiveness of this method is shown in equation(8), illustrating that although the probability of successful transmission between each repeater is high, the overall probability from Alice to Bob remains consistent as these events are independent. To address this, we simulate redundancy by conducting five simultaneous Bell state transmissions, enhancing entanglement swapping efficiency [42]. Fig.(3) depicts the use of quantum repeaters in entanglement swapping using an external Bell pair through intermediate repeaters.

Once entanglement swapping is performed and the trusted nodes are connected, they proceed with the QKD protocol as if the qubits have been successfully transmitted.

B. QKD Protocol Simulation

In our QKD simulation, transmissions are completed using either quantum repeaters or direct fiber-optic cables as outlined in section II. After conducting $10^5 - 10^6$ QKD rounds, we simulate error correction factoring in a decoherence value of $D = 0.02$ from environmental noise impacting all qubits. We identify erroneous qubits, calculate the error rate Q, and perform error corrections to derive the final key pool $K_{i,j}$ between trusted nodes using equation(9):

$$K_{i,j} = R(1 - 2h(Q_{i,j})), \tag{9}$$

where $h(Q)$ is the binary entropy function and R is the raw-keyrate for the protocol run. Trusted nodes exchange key material through XOR operations. For example, node T_2 assists in securely transmitting keys between Alice and Bob as shown in equation(10:

$$K_{A,2} = (K_{A,2} \oplus K_{2,B}) \oplus K_{2,B} \tag{10}$$

The overall key rate, which has the standard units of key-bits/s, indicating the QKD protocol's efficiency, is calculated by dividing the final key pool size by the number of QKD rounds.

C. Torus Topology

In this study, we present a $3D$ topology for our network simulation along with several traditionally used topologies. Fig.(2f) shows the torus topology generated through the network simulator whereas Fig.(4) shows the schematic representation of the $2D$ representation of torus topology, highlighting the connection between the user nodes. The idea behind torus topology is a $3D$ wrapping of the connections in a traditional grid topology.

IV. Results

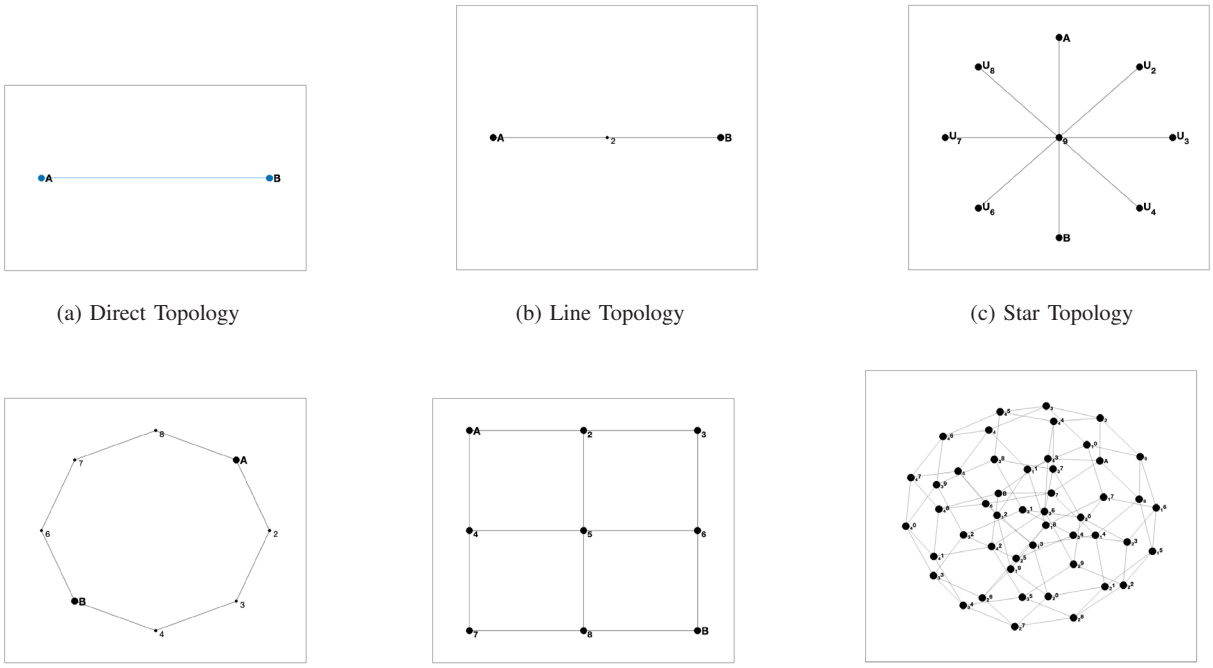
In this section, we'll go over the results obtained through our simulations for different QKD Protocols as mentioned in section II and over different Topologies.

A. Performance over Direct Topology

We first looked at the performance of different QKD protocols in direct topology because it serves as the most basic case of topology without any complexities. In this case, we simulated E91 in two cases, i.e., with and without using quantum repeaters. From Fig.(5), we can see that using quantum repeaters significantly increased the distance of stable

TABLE I
COMPARISON OF DIFFERENT TOPOLOGIES

Topology	Advantages	Disadvantages
Line	Offers one path between Alice and Bob, i.e., minimum control-layered overhead	• Impractical over long ranges. • Not reliable.
Gird	• Provides multiple path between Alice and Bob. • Nodes are geographically isolated.	Maximum control-layer overhead needed.
Ring	Provides two paths between Alice and Bob with lesser control-layer overhead.	Less reliable than grid topology.
Star	Allows several user nodes at once.	Reduces overall performances and increased risk.
Torus	Higher connectivity due to multiple available paths, and comparatively easier expansion without much reconstruction	Higher initial and overall maintenance due to complex structure architecture.



(a) Direct Topology (b) Line Topology (c) Star Topology

(d) Ring Topology (e) Grid Topology (f) 7×7 Torus Topology.

Fig. 2: Different topologies used in our QKD simulations (all of the topology profiles are generated through our simulations directly).

key rates to the case without any quantum repeaters. We see that three-stage protocol offers high key-rates for lower distances, however it reduces significantly with increasing distances. This can be associated with the fact that qubits have to travel thrice the distance than other protocols, thus attenuation loss is significantly higher. Decoy state offers the least key-rates due to vacuum bits occupying the bandwidth, and with similar attenuation, the number of effective qubits contributing to final key-rate is significantly lower. We see that E91 with repeater offers slightly lower key-rates, however it offers more stable key-rate over larger distances.

B. Performance of QKD Protocols over Different Topologies

This section explores the performance of different QKD protocols over different Topologies as described in Fig.(2). We compare the changing key rates over various distances between Alice and Bob for the 3-stage protocol, the E91 protocol, and decoy-state protocol (section II). As seen in Fig.(6), Decoy-State does not offer very stable transmission distances, and the decay is more rapid than the other two protocols. One other interesting thing to note is that the grid topology (as shown in Fig.(2e) offers significantly higher key rates for 3-stage protocol due to the availability of multiple paths for key distribution, i.e., the possibility of more than one key being distributed each round because of the presence of multiple trusted nodes.

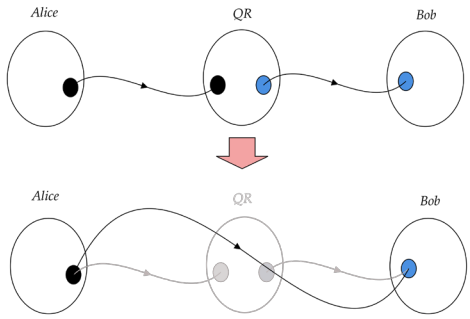


Fig. 3: Working of the quantum repeater: Starts with two sets of entangled bell-pair between Alice-QR and QR-Bob. Then, we perform an entanglement swap between the nodes to create a final entangled Bell state pair between Alice and Bob. The curly line represents an entangled pair.

TABLE II
SUMMARY OF QUANTUM COMMUNICATION PROTOCOL SIMULATION PARAMETERS.

Parameter	Description
α	Fiber attenuation coefficient. $\alpha = 0.15$. $\alpha = 0.4$ for optical switches.
L	Length of fiber segments in kilometers.
Burst Size	Frequency of photons in one burst. The default values are as follows, • Decoy State: A probability distribution between 0 and 2. • Three-Stage: 10 photons sent in the first burst, then subsequent numbers. • E91: 5 parallel Bell-state attempting between each quantum repeater.
B	Probability of successful Bell state measurements for each of the quantum repeaters in between Alice and Bob. $B = 0.85$.
D	Probability of quantum state decoherence due to channel noise. $D = 0.02$.

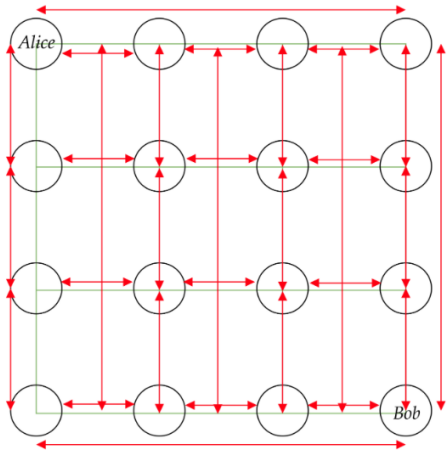


Fig. 4: Schematic representation of the Torus topology. Red lines highlights the connection logic between user nodes.

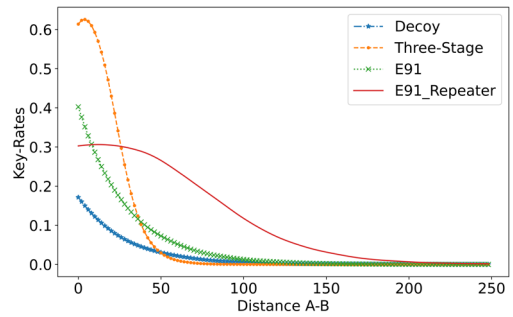


Fig. 5: Comparison of Performance of different QKD Proto- cols over Direct Topology.

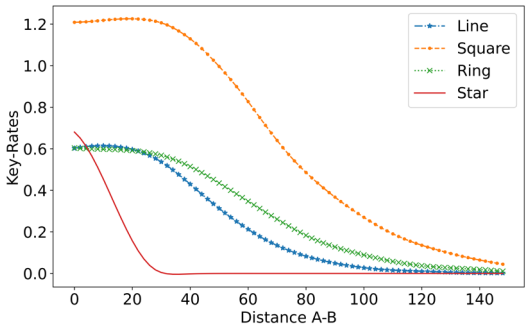
QKD Protocol Performance over Torus Topology: This section explores the performance of the 3-stage protocol and E91 protocol over our torus topology. As we can see from Fig.(2f), there are multiple paths between Alice and Bob, and unlike grid topology (Fig.(2e)), the shortest path between Alice and Bob is L itself. This increases the probability of successful transmission between Alice and Bob, and thus, we can expect a higher key rate than that of the grid topology. As shown earlier in Fig.(5) the key rates were observed to be higher for the 3-stage Protocol than the E91 Protocol. We see similar behavior from Fig.(10) that the 3-stage Protocol beats the E91 Protocol almost by 1.5 times. We also see that the stable distance for E91 is observed to be higher, as seen in the previous cases as well.

C. Performance of the Decoy-State Protocol

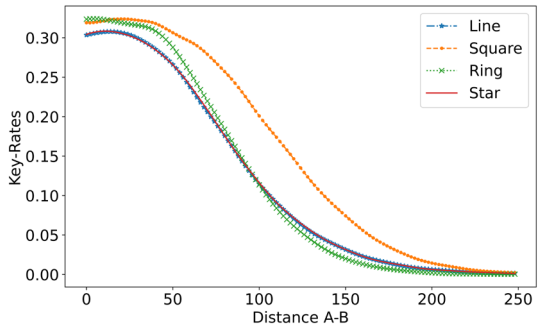
From Fig.(6c), we can see that the grid topology offers the highest key rates than the rest of the topologies due to multiple paths and multiple trusted nodes contributing to more than a single key per simulation round. The grid topology also seems to be more durable over time because the fiber segments between trusted nodes are shorter compared to other topologies used. Ring topology also offers multiple paths, but it seems to be just slightly better than line topology. However, over time, ring topology seems to show a better key rate with a lower decline rate than the line topology. For star topology, the higher attenuation coefficient ($\alpha = 0.4$ because of optical switches) results in significantly more fiber loss than any other topologies as described by equation(8).

D. Performance of the Three-Stage Protocol

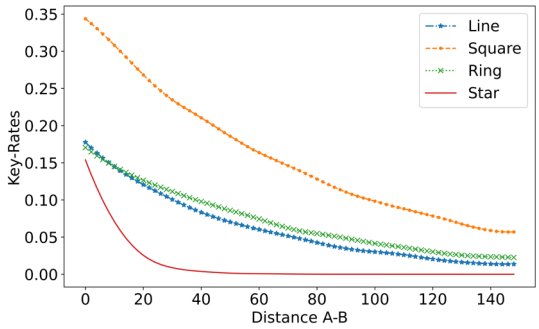
Figure 6a indicates that the grid topology consistently achieves higher key rates due to its multiple paths and trusted nodes, enhancing robustness and key pool size. Contrasting the 3-stage protocol's performance with direct and line topologies, as shown in figures Fig.(5) and Fig.(6a), the line topology exhibits slower key-rate decline over longer distances due to shorter fiber segments and a higher probability of successful transmission (equation(8). At shorter distances, line and ring topologies perform similarly, but the ring topology proves more durable at greater ranges, showing less key-rate decay.



(a) Three Stage Protocol Profile

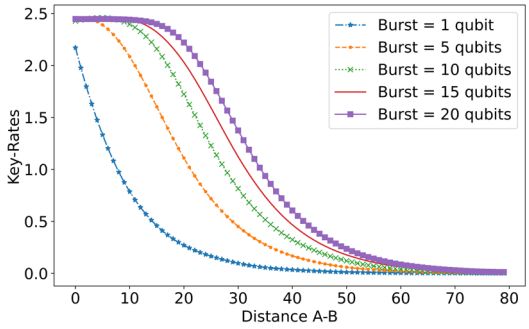


(b) E91 Protocol Profile

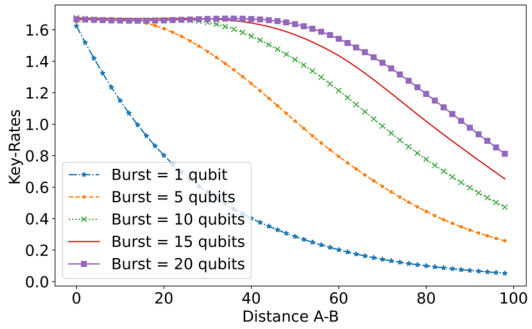


(c) Decoy-State Protocol Profile

Fig. 6: Performance of different QKD Protocols over Different Topologies



(a) 3- Stage QKD Protocol Performance over a 3×3 Torus Topology.



(b) E91 QKD Protocol Performance over a 3×3 Torus Topology.

Fig. 7: 3-Stage and E91 Protocol Performance over Torus Topology.

Conversely, the star topology, despite starting with higher key rates, experiences rapid decay because of increased attenuation from simulating an optical switch.

E. Performance of E91 Protocol

Fig.(6b) demonstrates that while most topologies show similar performance under the E91 protocol, the grid topology shows robustness and slower key rate decay, attributed to its multiple paths and three repeaters in the shortest path. The ring topology, in contrast, achieves higher key rates at very short distances due to fewer repeaters, reducing the likelihood of entanglement swapping failures. Despite its benefits, the grid topology's advantage is offset by errors from failed quantum repeaters. Conversely, the star topology leverages a central repeater to directly connect any two nodes by creating Bell states selectively, circumventing the need for an optical switch and matching the efficiency of direct line topologies that rely on a single quantum repeater for connectivity between nodes like Alice and Bob.

F. Analysing Higher Order Multi-Photon Bursts

In this section, we explore the trade-offs between performance and security risks with larger multi-photon bursts in the 3-stage protocol, which previously utilized a default burst size of 10 qubits. While increasing burst size can enhance performance, it also raises the potential for PNS attacks. For the E91 protocol, which does not naturally support multi-photon bursts, performance adjustments involve varying the number of simultaneous Bell pairs between quantum repeaters.

Fig.(8a) demonstrates that larger burst sizes generally maintain stable key rates at shorter distances for the 3-stage protocol. Meanwhile, Fig.(8b) shows that the E91 protocol sustains longer-range stable transmissions than the 3-stage protocol, as corroborated by Figures Fig.(5) and Fig.(6b). However, both protocols see key rates diminish at extended distances, with variations depending on the burst size and protocol. These findings highlight the importance of exploring various network topologies and integrating multiple trusted nodes to enhance protocol efficacy.

G. Multi-Photon Burst and Distance Relation

This section looks into finding a mathematical relation governing these curves to define a predictive model for the simulations. We notice a few features from the curves,

- A stable flat curve for smaller distances between Alice and Bob.
- A decaying curve, sort of linear, in the middle section.
- The curve converges to zero at larger distances.

We define an exponentially decaying function of the sigmoid nature as described in equation(11),

$$y = \frac{R}{(1 + e^{k(x-x_0)})}, \tag{11}$$

where, R, k, x_0 are the fitted parameters, and x is the inter-node distance between Alice and Bob (i.e., the sender and the receiver). R defines the initial constant value of key-rate, k defines the decaying rate of the curve, and x_0 is the point of the curve where key-rate $= R/2$. We do curve fitting for the line and ring topologies and present the results below in Fig.(9).

Based on the values of various fitted parameters from equation(11), the following were the equations of the curves found for the above are,

$$y_{\text{line}} = \frac{0.655}{(1 + e^{0.139(x-25.303)})} \text{ and } y_{\text{ring}} = \frac{0.652}{(1 + e^{0.109(x-32.257)})} \tag{14}$$

equation(12) describes the characteristic equations for line and ring topology for $3-$stage protocol.

1) Multi-Photon Burst Profiles: This section explores the multi-photon burst relations for higher order bursts over line topology for the 3-stage protocol. We explore the key rates and distance relation for the following burst sizes, $b = [50, 1200]$. We first make the curve smooth by eliminating the polynomial noise using the SavGol filter. We then fit the curves using equation(11) and present the characteristic curves for each of them in Fig.(10).

For lower order burst size, we see the fit of the curve being better as key rates approaches zero (for Fig.(10a)), and for higher order curves (such as Fig.(10b)) we can clearly see that the curve fit becomes better for the initial region as well. One important thing to note is that we can see the constant part of the curve increases with the increasing multi-photon burst size; therefore, we need to establish a relationship between the two quantities as well. This will be done in the next section.

H. Distance of Stable Transmission and Multi-Photon Burst Size

It is evident that the multi-photon burst increases the distance of stable transmission for all of the curves that we found. This is to be noted that this result can be generalized for same protocol at different multiphoton burst size. This can be done by using the following steps. To find this relation, we have to first find the *turning* point, i.e., the point where the derivative of the curve becomes negative.

From Fig.(11), we can see that a 3rd order polynomial fit describes the relationship very well. Therefore, we fit a third-order polynomial to the data points and find the characteristic equation for the curve. The equation of the third-order polynomial curve was found as shown in equation(15).

$$y = 0.000008x^3 - 0.003388x^2 + 0.538443x + 1.693613 \tag{15}$$

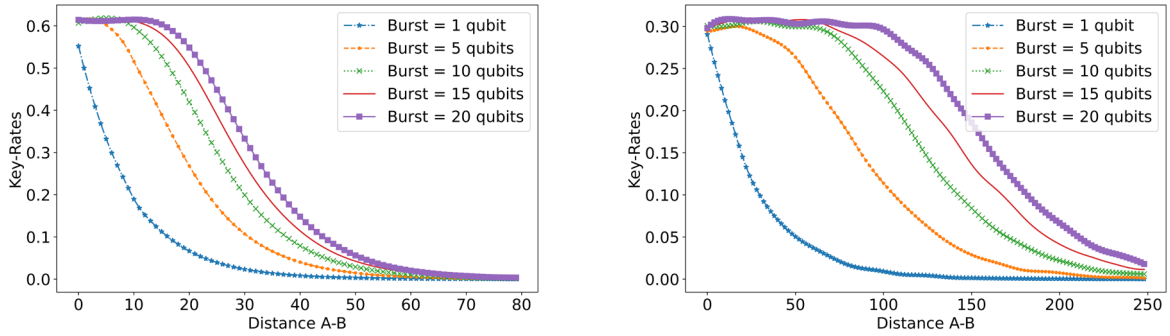
From equation(15), we can see that multi-photon burst size follows a polynomial relationship with the distance of stable transmissions. However, the burst sizes used in this calculation are of very small order. To get a general picture and a more practical burst-size relation, we increase the burst size up to a million qubits at once. Fig.(12) shows the curve with stable transmission distance and multiphoton burst size for bursts of up to one million qubits at once.

Fig.(12) gives an insight into a generic relationship between the higher-order bursts and the distance of stable transmission. We set a curve between $\log(\text{burst\_values})$ and the distance of stable transmission. Fig.(13) gives us a better picture of a possible relationship between the main two quantities of interest of this study, that is, the distance of stable transmission and the size of multiphoton burst used.

From Fig.(13), we found that the two quantities of interest share a logarithmic-3rd order polynomial relationship as described by equation(16). The model explains over 99% of the variance ($R^2 = 0.998$) and remains high after adjusting for its four parameters (adjusted $R^2 = 0.997$), showing that a cubic fit in $\log(\text{burst})$ works well. Furthermore, a confidence band of 95% also shows very less uncertainty in the fit.

$$D_s = -0.054x^3 + 1.228x^2 + 1.1878x + 2.0178, \tag{16}$$

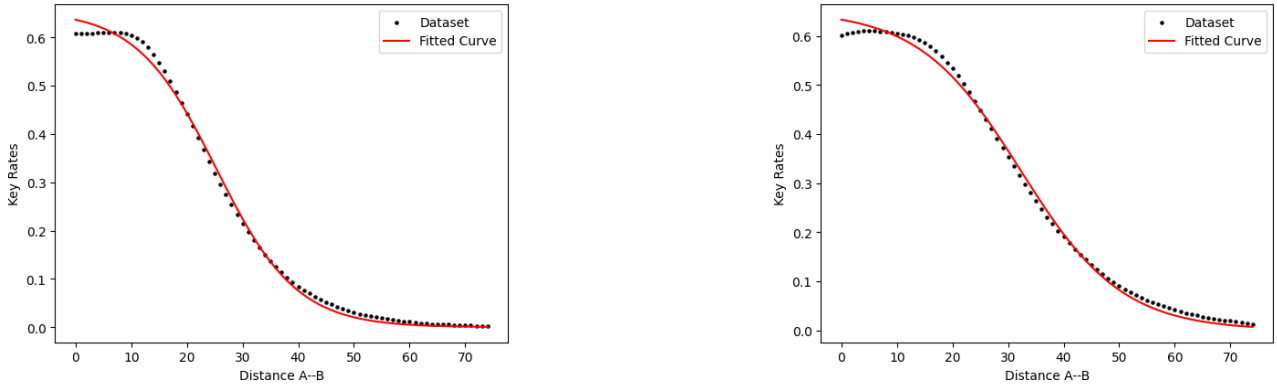
where $x = \log(\text{burst\_size})$. Therefore, based on the above, we can define a generic relationship for a 3-stage protocol over a line topology between the distance of stable transmission (denoted by D_s), and the size of multi-photon burst used as described in equation(17).



(a) Multi-Photon Burst Profile for Three-Stage Protocol over Line Topology.

(b) Multi-Photon Burst Profile for E91 over Line Topology.

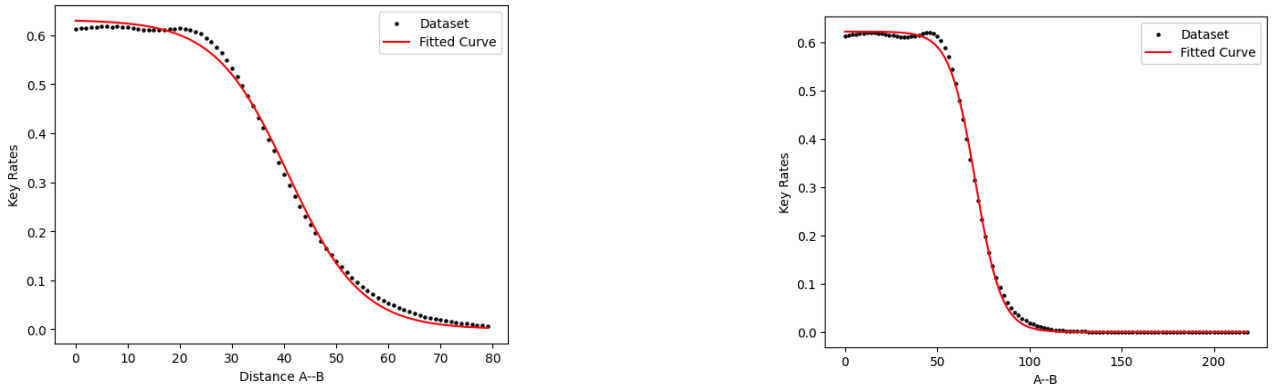
Fig. 8: Comparison of Multi-Photon Burst Size Profiles for Three-stage and E91 Protocol



(a) Characteristic curve obtained for QKD Performance over a Line-topology.

(b) Characteristic curve obtained for QKD Performance over a Ring-topology.

Fig. 9: Fitting the curve to the equation(11) for Line and Ring topology for a multi-photon burst size of 10 qubits for 3-stage QKD Protocol.



(a) Characteristic curve obtained for burst size $= 50$ qubits for 3-stage protocol.

(b) Characteristic curve obtained for burst size $= 1200$ qubits for 3-stage protocol.

Fig. 10: Fitting the curve to the equation(11) for line topology having higher order multi-photon bursts used for 3-stage Protocol.

$$D_s = -\phi \log^3(b) + \beta \log^2(b) + \gamma \log(b) + \delta, \qquad (17)$$

where, $\phi, \beta, \gamma,$ and δ are the curve fitting parameters and b is the size of the multi-photon burst used. This gives us insight into a rather counter-intuitive relationship occurring between the two quantities. For line topology with $B = 0.85$ and $D = 0.02$, and $\alpha = 0.15$, equation(16) can be used to determine the maximum distance of stable transmission for a given value of the multi-photon burst.

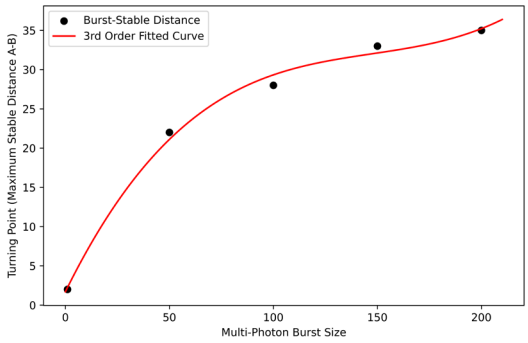


Fig. 11: Relationship between the Maximum Stable Distance (A-B) and the size of multi-photon burst used.

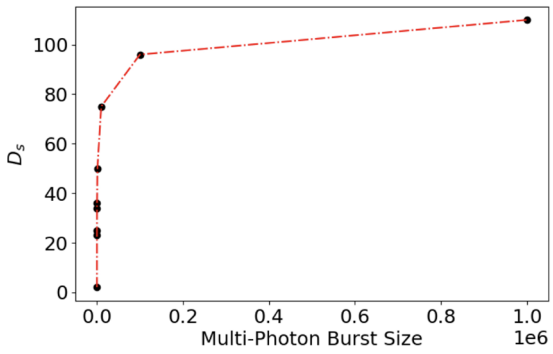


Fig. 12: Relationship between the maximum stable distance (between Alice and Bob) and the size of multi-photon burst used consisting of burst sizes up to 106 qubits. The dashed lines is just for visualization purposes, and to study the data- trend more effectively, we do curve fitting in log scale in next figure.

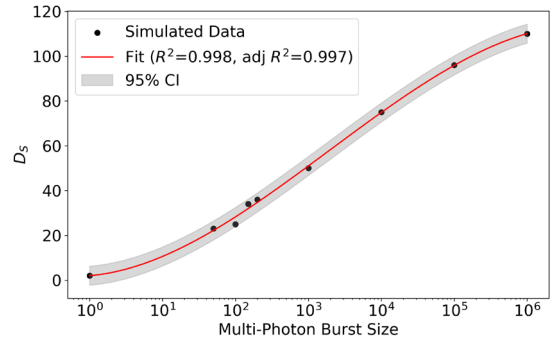


Fig. 13: Relationship between the Maximum Stable Distance (A-B) and the size of multi-photon burst used consisting of burst sizes up to 106 qubits (x-axis is in log scale). CI refers to the confidence index of the curve fitting.

V. Conclusions and Future Work

Quantum networks are increasingly becoming viable, as evidenced by empirical studies from initiatives such as the DARPA, European, and Tokyo quantum networks. These studies have addressed issues such as the distance between nodes and the stabilization of transmissions through the use of trusted nodes and optical switches. As indicated in equation(8), the probability of successful transmission is dependent on the attenuation coefficient and the distance between the nodes, with different optical materials exhibiting variable attenuation constants. This requires exploring diverse topologies to optimize node placement and attenuation impacts. Our study evaluates the performance of three prominent QKD protocols—Decoy-state, 3-stage, and E91—across various network topologies.

Our findings indicate that the grid topology outperforms simpler configurations by leveraging multiple paths and trusted nodes, enhancing the key pool and robustness across the three explored QKD protocols. Notably, when we analyzed the torus topology with the 3-stage and E91 protocols, it yielded significantly higher key rates than any other topology. Additionally, we derived a mathematical model to quantify key-rate variations with increasing distances between Alice and Bob over the line topology. We also formulated an equation for the maximum feasible distance for stable transmission in a line topology with three nodes, laying foundational insights for practical quantum network designs. This advancement underscores the importance of tailoring the multi-photon burst size to the node separation, optimizing communication efficacy in practical quantum networks. Apart from these, the multiphoton approach studied in this work can be used in multi-carrier continuous-variable QKD (CVQKD) systems, where coherent states are multiplexed into Gaussian subcarrier channels to increase aggregate key rates and resilience against channel noise[43]. All of the results presented in this work define future frameworks for a more scalable and reliable network, laying foundation for the development of Quantum Internet[44].

As future work, we aim to establish a relationship between multi-photon burst sizes and maximum stable transmission distances across various topologies, enhancing our understanding of multi-photon device behaviors. There is also a need to address practical challenges associated with current QKD protocols, which often rely on heuristics and may not provide optimal solutions, thereby limiting their practicality for larger network development [45]. Moreover, in the NISQ era, network engineering must account for Johnson noise and other real-world phenomena, necessitating the development of advanced error-correction methods to improve quantum router performance in noisy environments [46]. A proposed solution to these problems in recent times has been quantum augmented networks. In this approach, several quantum components (such as quantum teleportation[47], QKD, quantum error correction, etc.) are strategically integrated in current classical networks to enable large-scale quantum communication [48], [49].

In summary, this study not only highlighted the advantages of multi-photon QKD protocols across different topologies but also established an empirical relationship between burst size and stable transmission distances for the 3-stage protocol. This protocol demonstrated enhanced capabilities for transmitting higher-order qubit bursts, yielding more stable and higher key rates. While the empirical relationship was specifically developed for the 3-stage protocol over the line topology, the insights gained will guide the selection of suitable protocols for crafting more robust quantum networks.

ACKNOWLEDGMENT

This work is sponsored by NSF award #2324924 and #2324925.

REFERENCES

[1] Z. Zhang and Q. Zhuang, "Distributed quantum sensing," *Quantum Science and Technology*, vol. 6, no. 4, p. 043 001, 2021. DOI: 10.1088/2058-9565/abd4c3.

[2] N. Aslam, H. Zhou, E. K. Urbach, et al., "Quantum sensors for biomedical applications," *Nature Reviews Physics*, vol. 5, no. 3, pp. 157–169, 2023. DOI: 10.1038/s42254-023-00558-3.

[3] P. Zhang, N. Chen, S. Shen, S. Yu, S. Wu, and N. Kumar, "Future quantum communications and networking: A review and vision," *IEEE Wireless Communications*, 2022. DOI: 10.1109/MWC.012.2200295.

[4] R. Renner, "Security of quantum key distribution," *International Journal of Quantum Information*, vol. 6, no. 01, pp. 1–127, 2008. DOI: 10.1142/S0219749908003256.

[5] National Security Agency/Central Security Service, *Quantum key distribution (qkd) and quantum cryptography qc*, https://www.nsa.gov/Cybersecurity/Quantum-Key-Distribution-QKD-and-Quantum-Cryptography-QC/, Accessed: [16 December 2023], 2023.

[6] C. Elliott, "Building the quantum network," *New Journal of Physics*, vol. 4, pp. 46.1–46.12, 2002. DOI: 10.1088/1367-2630/4/1/346.

[7] B. Yurke and J. S. Denker, "Quantum network theory," *Journal Title*, vol. 29, no. 3, Page Range, Mar. 1984. DOI: 10.1103/PhysRevA.29.1419.

[8] T. Satoh, S. Nagayama, S. Suzuki, T. Matsuo, M. Hajdusek, and R. V. Meter, "Attacking the quantum internet," *IEEE Transactions on Quantum Engineering*, vol. 2, pp. 1–17, 2021. DOI: 10.1109/TQE.2021.3094983.

[9] M. Peev, C. Pacher, R. Alléaume, et al., "The secoqc quantum key distribution network in vienna," *New Journal of Physics*, vol. 11, p. 075001, 2009. DOI: 10.1088/1367-2630/11/7/ 075001.

[10] M. Sasaki, M. Fujiwara, H. Ishizuka, et al., "Field test of quantum key distribution in the tokyo qkd network," *Optics Express*, vol. 19, p. 10 387, 2011. DOI: 10.1364/OE.19.010387.

[11] D. Ribezzo, M. Zahidy, I. Vagniluca, et al., "Deploying an inter-european quantum network," *Advanced Quantum Technologies*, 2023. DOI: 10.1002/qute.202200061.

[12] S.-K. Liao, W.-Q. Cai, W.-Y. Liu, et al., "Satellite-to-ground quantum key distribution," *Nature*, vol. 549, pp. 43–47, 2017. DOI: 10.1038/nature23655.

[13] S. Wehner, D. Elkouss, and R. Hanson, "Quantum internet: A vision for the road ahead," *Science*, vol. 362, eaam9288, Oct. 2018. DOI: 10.1126/science.aam9288.

[14] D. Unruh, "Everlasting multi-party computation," in *Advances in Cryptology – CRYPTO 2013*, R. Canetti and J. A. Garay, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 380–397. DOI: 10.1007/978-3-642-40084-1_22.

[15] A. A. Gaidash, V. I. Egorov, and A. V. Gleim, "Revealing of photon-number splitting attack on quantum key distribution system by photon-number resolving devices," in *Journal of Physics: Conference Series, IOP Publishing*, vol. 735, St.Petersburg, 2016, p. 012072. DOI: 10.1088/1742-6596/735/1/012072.

[16] C. H. Bennett, "Quantum cryptography using any two nonorthogonal states," *Physical Review Letters*, vol. 68, pp. 3121–3124, May 1992. DOI: 10.1103/PhysRevLett.68.3121.

[17] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Theoretical Computer Science*, vol. 560, pp. 7–11, 2014. DOI: 10.1016/j.tcs.2014.05.025.

[18] A. K. Ekert, "Quantum cryptography based on bell's theorem," *Physical Review Letters*, vol. 67, pp. 661–663, Aug. 1991. DOI: 10.1103/PhysRevLett.67.661.

[19] E. Diamanti, H. Lo, B. Qi, et al., "Practical challenges in quantum key distribution," npj *Quantum Information*, vol. 2, 2016. DOI: 10.1038/npjqi.2016.25.

[20] S. Sajeed, P. Chaiwongkhot, A. Huang, et al., "An approach for security evaluation and certification of a complete quantum communication system," *Scientific Reports*, vol. 11, no. 1, p. 5110, Mar. 2021. DOI: 10.1038/s41598-021-84139-3.

[21] P. K. Verma, E. M. Rifai, and C. K. W. Clifford, "Photon-number splitting attack," in *Multi-photon quantum secure communication*, Springer Nature Singapore Pte Ltd.: Springer, 2019, ch. 3. DOI: 10.1007/978-981-10-8618-2.

[22] B. Hensen et al., "Experimental loophole-free violation of a bell inequality using entangled electron spins separated by 1.3 km," *Nature*, vol. 526, p. 682, 2015. DOI: 10.1038/nature15759.

[23] T.-Y. Chen, X. Jiang, S.-B. Tang, et al., "Implementation of a 46-node quantum metropolitan area network," npj *Quantum Information*, vol. 7, p. 134, Dec. 2021. DOI: 10.1038/s41534-021-00474-3.

[24] S. Kak, "A three-stage quantum cryptography protocol," *Foundations of Physics Letters*, vol. 19, no. 3, pp. 293–296, Jun. 2006. DOI: 10.1007/s10702-006-0520-9.

[25] N. Jha, A. Parakh, and M. Subramaniam, "Effect of noise and topologies on multi-photon quantum protocols," in *Quantum Computing, Communication, and Simulation IV*, SPIE, vol. 12911, 2024, pp. 148–161. DOI: 10.1117/12.3000586.

[26] A. Parakh and M. Subramaniam, "Bootstrapped qkd: Improving key rate and multi-photon resistance," in *Quantum Information Science and Technology IV*, SPIE, vol. 10803, 2018, pp. 22–27. DOI: 10.1117/12.2500438.

[27] K. Thapliyal and A. Pathak, "Kak's three-stage protocol of secure quantum communication revisited: Hitherto unknown strengths and weaknesses of the protocol," *Quantum Information Processing*, vol. 17, no. 9, p. 229, Jul. 2018. DOI: 10.1007/s11128-018-2001-z.

[28] N. Jha, A. Parakh, and M. Subramaniam, "Joint encryption and error correction for secure quantum communication," *Scientific Reports*, vol. 14, no. 1, p. 24542, 2024. DOI: 10.1038/s41598-024-75212-8.

[29] J. Teng, Z.-Q. Yin, G.-J. Fan-Yuan, et al., "Sending-or-not-sending twin-field quantum key distribution with multiphoton states," *Physical Review A*, vol. 104, no. 6, p. 062 441, 2021. DOI: 10.1103/PhysRevA.104.062441.

[30] A. Biswas, A. Banerji, N. Lal, P. Chandravanshi, R. Kumar, and R. P. Singh, "Quantum key distribution with multiphoton pulses: An advantage," *Optics Continuum*, vol. 1, no. 1, pp. 68–79, 2022. DOI: 10.1364/OPTCON.445727.

[31] A. Mihály and L. Bacsárdi, "Effects of selected noises on the quantum memory satellite based quantum repeaters," *Infocommunications Journal*, vol. 13, no. 2, pp. 19–24, 2021. DOI: 10.36244/ICJ.2021.2.3.10

[32] T. Bisztray and L. Bacsardi, "The evolution of free- space quantum key distribution," *Infocommunications Journal*, vol. X, no. 1, pp. 22–30, Mar. 2018. DOI: 10.36244/ICJ.2018.1.4.

[33] N. Jha, A. Parakh, and M. Subramaniam, "Multi-photon 3-stage qkd for practical quantum networks," availavle at https://www.researchsquare.com/article/rs-3826628/v1, 2024. DOI: 10.21203/rs.3.rs-3826628/v1.

[34] V. Scarani et al., "The security of practical quantum key distribution," *Reviews of Modern Physics*, vol. 81, p. 1301, 2009. DOI: 10.1103/RevModPhys.81.1301.

[35] C. C. W. Lim, M. Curty, N. Walenta, F. Xu, and H. Zbinden, "Concise security bounds for practical decoy-state quantum key distribution," *Physical Review A*, vol. 89, p. 022 307, Feb. 2014. DOI: 10.1103/PhysRevA.89.022307.

[36] T. Attema, J. W. Bosman, and N. M. P. Neumann, "Optimizing the decoy-state bb84 qkd protocol parameters," *Quantum Information Processing*, vol. 20, p. 154, Apr. 2021. DOI: 10.1007/s11128-021-03078-0.

[37] Y. Zhao, B. Qi, X. Ma, H.-k. Lo, and L. Qian, "Simulation and implementation of decoy state quantum key distribution over 60km telecom fiber," in *2006 IEEE International Symposium on Information Theory*, Jul. 2006, pp. 2094–2098. DOI: 10.1109/ISIT.2006.261920.

[38] J. Burr, A. Parakh, and M. Subramaniam, "Evaluating different topologies for multi-photon quantum key distribution," in *Quantum Information Science, Sensing, and Computation XIV*, E. Donkor, M. Hayduk, M. R. Frey, S. J. L. Jr., and J. M. Myers, Eds., International Society for Optics and Photonics, vol. 12093, Orlando, Florida, United States: SPIE, 2022, p. 1 209 309. DOI: 10.1117/12.2620057.

[39] S. Mandal, G. Macdonald, M. El Rifai, et al., "Multi-photon implementation of three-stage quantum cryptography protocol," in *The International Conference on Information Networking 2013 (ICOIN)*, IEEE, 2013, pp. 6–11. DOI: 10.1109/ICOIN.2013.6496343.

[40] S. Felix, N. Gisin, A. Stefanov, and H. Zbinden, "Faint laser quantum key distribution: Eavesdropping exploiting multiphoton pulses," version 1, Feb. 2001. DOI: 10.1080/09500340108240903.

[41] O. Amer, W. O. Krawec, and B. Wang, "Efficient routing for quantum key distribution networks," in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, Denver, CO, USA: IEEE, Oct. 2020, pp. 137–147. DOI: 10.1109/QCE49297.2020.00027.

[42] W. J. Munro, K. A. Harrison, A. M. Stephens, S. J. Devitt, and K. Nemoto, "From quantum multiplexing to high-performance quantum networking," *Nature Photonics*, vol. 4, pp. 792–796, Nov. 2010. DOI: 10.1038/nphoton.2010.213.

[43] L. Gyongyosi, "Multicarrier continuous-variable quantum key distribution," *Theoretical Computer Science*, vol. 816, pp. 67–95, 2020. https://doi.org/10.1016/j.tcs.2019.11.026.

[44] J. Burr, A. Parakh, and M. Subramaniam, "Quantum internet," *Ubiquity*, vol. 2022, no. August, pp. 1–14, 2022. DOI: 10.1145/3547493.

[45] Y. Zeng, J. Zhang, J. Liu, Z. Liu, and Y. Yang, "Entanglement management through swapping over quantum internets," *IEEE Network*, 2023. DOI: 10.1145/3626570.3626595.

[46] W. Shi and R. Malaney, "Quantum routing for emerging quantum networks," *IEEE Network*, 2023. DOI: 10.1109/MNET.2023.3317821.

[47] A. Parakh, "Quantum teleportation with one classical bit," *Scientific reports*, vol. 12, no. 1, p. 3392, 2022. DOI: 10.1038/s41598-022-06853-w.

[48] N. Jha, A. Parakh, and M. Subramaniam, "Towards a quantum-classical augmented network," in *Quantum Computing, Communication, and Simulation V*, SPIE, vol. 13391, 2025, pp. 72–86. DOI: 10.1117/12.3042150.

[49] N. Jha, A. Parakh, and M. Subramanian, "A ml based approach to quantum augmented http protocol," in *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, IEEE, vol. 2, 2024, pp. 591–592. DOI: 10.1109/QCE60285.2024.10420.

Nitin Jha received his BSc. (Hons) in Physics from Ashoka University in 2023. He is pursuing his Ph.D. from Kennesaw State University (KSU). His broad research interests lie in quantum networks, secure quantum communication, and network security. He is looking at application of artificial intelligence techniques to improve the efficiency of quantum networks and building the foundations of quantum-classical hybrid networks (also known as quantum augmented networks). Nitin's research has won multiple accolades at KSU, appeared in top venues, and continues to be a part of the strategic initiative at the University.

Abhishek Parakh is a Professor of Computer Science and the Director of the Computer Science Ph.D. Program at Kennesaw State University. He received his Ph.D. in Computer Science from Oklahoma State University in 2011. He has held various academic positions, including Director of NebraskaCYBER and the Mutual of Omaha Distinguished Chair of Information Science and Technology. His research interests include quantum cryptography, cybersecurity, and trustworthy computing. He has been the principal investigator on several federally funded projects (NSF, DOD, NASA, NSA, DOS) related to quantum networks, quantum computing, post-quantum cryptography, AI driven educational technology, and cybersecurity education.

Mahadevan Subramaniam is the Charles W. and Margre H. Durham Distinguished Professor and Chair of the Computer Science Department at the University of Nebraska Omaha (UNO). He earned his B.S. in Computer Science from BITS Pilani in 1986 and his M.S. and Ph.D. in Computer Science from SUNY Albany in 1997. His research focuses on formal methods for software engineering, particularly model checking, SMT solvers, and theorem provers, with applications in software evolution and automated repair strategies.

Mode Selection in Mode Division Multiple Access System for In Building Solution in Mobile Networks

Ahmed S. Mohamed[1], and Eszter Udvary[1, 2]

Abstract—**This paper introduces the application of Mode Division Multiple Access (MDMA) in the context of In-Building Solutions (IBS) for mobile networks. The study showcases the successful generation and selection of light modes which are then efficiently multiplexed and demultiplexed at the Remote Radio Unit (RRU) end. Despite the proven operational capabilities, the findings reveal a decline in signal quality as the distance increases, thus limiting the use of MDMA for long-distance fronthaul applications. The proposed system also simplifies the RRU by centralizing key functionalities at the Central Office (CO), potentially reducing costs and the operational expenses (OPEX/CAPEX) associated with in- building solutions and other mobile network deployments. This work extends previous research and paves the way for future studies, particularly in the application of the Power over Fiber (PWoF) approach to reduce RRU complexities further.**

Index Terms—**In-Building Solutions, Mode Division Multiplexing, Spatial Laser Modulator**

I . INTRODUCTION

As LTE networks become increasingly common, there's a growing interest in Centralized Radio Access Networks, or CRANs. These networks are appealing because they can significantly lower both operational and initial costs. In a CRAN configuration, the system is divided into the baseband unit (BBU) and the remote radio unit (RRU). Essentially, a centralized group of BBUs handles the intense baseband processing and manages control and oversight tasks for numerous RRUs. The connection between the RRUs and BBUs, known as the fronthaul, depends on the well-established Common Public Radio Interface (CPRI) standard to work effectively.[1].

To address the complex needs of 5G and 6G networks and their advanced antenna systems, experts are increasingly considering Wavelength Division Multiplexing (WDM) as a solution for transmitting CPRI data across fronthaul networks. WDM offers several advantages: it enhances network efficiency, supports various types of data seamlessly, and helps save energy. However, the main challenge is making this technology affordable and flexible, especially in relation to the remote radio unit (RRU) components of the network[1]. The expense tied to the lasers used in RRUs poses a significant challenge to the commercial feasibility of this technology[1].

[1] Department of Networked Systems and Services, Faculty of Electrical Engineering and Informatics Budapest University of Technology and Economics, Budapest, Hungary

[2] ELKH-BME Information Systems Research Group, Budapest, Hungary (e-mail: asayed@hit.bme.hu)

DOI: 10.36244/ICJ.2025.2.10

As the call for more transmission capacity grows, Mode-Division Multiplexing (MDM) is also gaining traction. MDM is being scrutinized for its capacity to manage swift optical transmissions and enhance access networks, potentially serving an expanding number of users [2][3], [4]. Notably, MDM systems are not restricted by wavelength demands, which is a major plus for network services that require non-specific wavelength capabilities. This characteristic could lead to substantial cost cuts for the RRUs[1]. Moreover, the multiple access technique plays a vital role for access networks. In classical mobile communication networks, various multiple access techniques have been standard, such as Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), Code Division Multiple Access (CDMA), and Orthogonal Frequency-Division Multiple Access (OFDMA). FDMA works by splitting the available bandwidth into several distinct frequency bands, each assigned to an individual user. TDMA, on the other hand, segments time into separate frames, which are further divided into slots, with each time slot allocated to a different user to enable multiple users to share the same frequency channel[5]. CDMA employs spread spectrum technology, where a pseudo-random code modulates the data signal, widening its bandwidth before it's transmitted over a carrier wave[6]. However, these traditional multiple access methods have limitations in terms of resource allocation and are unable to cater to the massive connectivity and high-volume requirements of 5G networks. As a result, there's a pressing need for innovative multiple access strategies that can fulfill the demands of next-generation mobile communications, which emphasize supporting a large number of concurrent users and connections[6]. Frequency, time, code, and space are all dimensions that can independently separate users in communication systems, offering distinct degrees of freedom for multiple access methods. For instance, CDMA exploits frequency, time, and code in conjunction to differentiate between users, giving it three degrees of freedom. In contrast to these traditional dimensions, optical modes present an additional degree of freedom for millimeter-wave communications[5][7].

Despite the fact that the traditional multiple access techniques mentioned above provide the air interface to the end users devices in mobile networks, this paper introduces and explores a different multiple access strategy for millimeter-wave communication systems, a strategy that leverages the orthogonality and high dimensionality of optical modes where the access units here are the RRUs. The main contribution presented herein is the proposed adaptation of

contribution presented herein is the proposed adaptation of the Mode Division Multiple Access (MDMA) scheme, which utilizes the spatial characteristics of optical modes for multi-antenna millimeter-wave communication systems based on Mode Division Multiplexing (MDM). It is demonstrated that the use of optical modes as a new degree of freedom has the potential to significantly enhance the capacity and spectral efficiency of communication systems, with a particular advantage for In-Building Solutions (IBS). Moreover, this paper shows that a bidirectional MDM transmission system that centralizes the generation and excitation of modes at the point of transmission origin can streamline the design of Remote Radio Units (RRUs), potentially making them less complex and more cost-effective to produce[1].

II. SYSTEM CONCEPT

MDM is recognized as a potent method that has the potential to vastly expand transmission capacities within high-speed optical data transport and access networks, thereby enabling the support of considerably expanded user bases[2] . MDM technique is distinct from conventional methods that use single-mode fibers (SMF), which transmit a single data stream, and multimode fibers (MMF), which have less control over the transmitted signals. MDM utilizes the orthogonal properties of light modes to create multiple independent data channels within a single fiber, offering a stark contrast to wavelength-based multiplexing techniques such as Coarse and Dense Wavelength Division Multiplexing (CWDM and DWDM), which segregate channels by different wavelengths of light. [1]. While multimode fibers (MMFs) have the theoretical capability to support numerous modes and thereby enhance capacity, their practical application over extended distances is limited due to issues like crosstalk, where there is unwanted energy transfer between modes, and modal dispersion, which can cause signal distortion[8] . These issues compromise the integrity of the signal, which can result in a high Bit Error Rate (BER). Consequently, this substantially restricts the bandwidth-distance product, a measure of the data-carrying capacity of the fiber over a given distance. [8]. To address these challenges, Few Mode Fibers (FMFs) have been introduced. They are engineered with a core size that is intermediate between that of Single Mode Fibers (SMFs) and Multimode Fibers (MMFs), allowing FMFs to support a finite, manageable number of modes. This innovation curtails crosstalk and improves signal clarity, rendering FMFs a more viable option for long-distance communication[9].

Expanding on MDM, MDMA further improves network adaptability. In an MDMA framework, distinct modes are allocated to different remote devices RRUs. This method sets up separate transmission paths within the same optical fiber, thus achieving a level of physical layer isolation. [10]. A solitary FMF linking the Central Office (CO) to various Remote Radio Units (RRUs) is capable of sustaining numerous dedicated channels. This configuration presents an effective IBS, equipped with ample capacity to cater to multiple RRUs simultaneously.

Although MDM and MDMA offer considerable benefits, its full-scale commercial deployment encounters various hurdles. The technologies for effectively multiplexing and demultiplexing light modes are still in the developmental phase and require additional enhancements to achieve peak functionality [1]. Furthermore, to address issues such as crosstalk and modal dispersion, there is often a need to implement sophisticated Digital Signal Processing (DSP) methods. While these DSP techniques are potent, they have the potential to add to the system's intricacy [11][12].

A bidirectional and symmetrical MDM system with MDMA as the multiple access technique that is designed for high-speed mobile fronthaul applications, is proposed. For simplicity, the system under consideration involves a single CO and two RRUs, necessitating the use of four spatial modes, two modes for the downlinks and the other for the uplinks. The transmission of the downlink RF signal is facilitated by the first two modes, while the transmission of an unmodulated carrier from the Baseband Unit (BBU) pool is carried out by the second two modes. At the RRU, the carrier is received and modulated with the uplink RF signal. This method centralizes mode generation, thereby reducing the complexity and cost of the RRU by obviating the need for costly laser components. Crosstalk issues, often encountered in wavelength-reuse schemes, are minimized by this design, which enables symmetrical bidirectional transmission [1]. Figure 1 shows the system's concept structure, featuring the paths for the carrier, the RF/Downlinks, and the RF/Uplinks. At the CO, four optical modes are generated and sent to the RRU through an FMF. The modes designated for downlinks are then picked up at the RRU, where they are filtered, and the Radio Frequency (RF) signals are dispatched to their respective antennas. Conversely, the RF signals for uplinks, collected by the antennas, undergo filtering and amplification before modulating the latter two optical modes, which initially act as carriers. These modulated modes are subsequently merged back into the FMF via one or more circulators back to the CO.

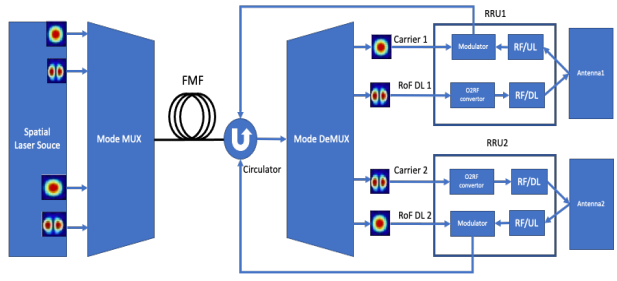


Fig. 1. System Concept

Beyond centralizing mode generation to simplify the RRU design, incorporating a PWoF solution could offer additional advantages. By delivering both power and data through the same optical fiber, PWoF eliminates the need for separate electrical infrastructure at the RRU site [13]. This approach is particularly valuable in scenarios where space or access to power is limited, such as in-building or remote installations. As this technology continues to mature, integrating PWoF into MDMA-based systems could lead to more compact, energy-efficient, and cost-effective fronthaul architectures [14].

III. SIMULATION AND MODELLING

The modeling of the system was carried out on 400m FMF utilizing the VPI Photonics simulation environment. This platform is esteemed for its extensive use in crafting and examining designs in the fields of integrated photonics, fiber optics, optoelectronic components, and optical transmission systems. It served as the principal tool in simulating the intricate details of our photonic system.

A. FMF Modelling

In MDM systems that utilize FMFs, unlike in conventional multimode fibers (MMFs) where mode excitation is a spontaneous process, a high degree of precision is required to govern the excitation of specific types and numbers of modes. This careful control over mode excitation is critical to achieve optimal transmission properties before the modes are coupled into the FMF. [1][15]. As previously stated, our proposal includes the use of an FMF that is designed to support four modes. The selection of the FMF's core diameter was determined by applying the natural frequency formula, considering the refractive indices of both the core and the cladding materials that are found in commercially available fibers [1] [16]. This analysis led to the selection of a fiber with a core diameter of 20 µm, which is not commercially available but can be ordered for customized research purposes like this study [17]. Various mode selection techniques have been developed by researchers to establish this level of control. [18]–[21]. The parameters of the Few Mode Fiber (FMF) utilized in the simulation were meticulously defined, drawing upon the mathematical calculations expounded in the preceding section. The excitation of the designated number of modes within the simulated FMF was confirmed. As shown in Figure 2, the simulation results affirm the precision of our theoretical model, demonstrating that a 20µm fiber supports four modes.

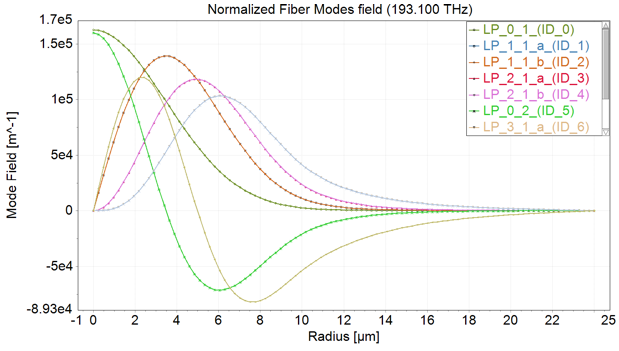


Fig. 2. Excited modes

B. Mux/DeMux

The task of effectively multiplexing and demultiplexing light modes continues to be a dynamic field of study. [1]. Photonic lanterns are especially distinguished among the various strategies suggested for multiplexing and demultiplexing in MDM due to their structural simplicity. [22]. In this method, a specific quantity of Single Mode Fibers (SMFs), each featuring unique core and cladding dimensions, are fine-tuned to align with the required mode number. These altered SMFs are collectively encased within a single sheath and undergo a meticulous tapering process, induced by controlled heating, to meet the adiabatic criterion, thus creating a photonic lantern. Nonetheless, the slimmed-down core of the tapered SMF makes the light field of the guided mode more prone to escaping into the cladding. Given that the refractive index of the quartz sleeve is less than that of the SMF cladding, a novel waveguide structure emerges between the cladding of the tapered fiber and the quartz sleeve [22]. Figure 3 presents the conceptual design of photonic lanterns. In this configuration, a component with reciprocal functionality is utilized to multiplex (MUX) and demultiplex (DeMUX) the modes into and out of the FMF, with its parameters appropriately adjusted for the process.

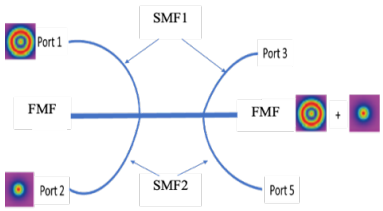


Fig. 3. Photonic Lanterns [22]

Within the simulation environment, the corresponding component named “CombinerSplitterMM” was utilized to emulate the photonic lanterns and to multiplex/demultiplex the modes within the FMF.

C. Modes Selection

In this configuration, we use a leading simulation software called VPI Photonics, which specializes in photonics design across integrated photonics, fiber optics, optoelectronics, and optical transmission systems. Spatial Light Modulators (SLM) are one of the simplest approaches used in modes selection where phase plates that have phase profiles match the targeted mode are used[15], [21].

The setup of this study consists of using laser sources equivalent to the number of modes. As mentioned in the previous section, a component named “CombinerSplitterMM” was utilized to emulate the photonic lanterns and to multiplex/demultiplex the modes within the FMF. To simulate SLM in the simulator, an equivalent component named “CouplerBeamMM” is used as the setup’s SLM. By setting the suitable parameters, it was possible to select the higher order modes LP11a, LP21a and LP31a. On the other hand, the fourth mode of choice is LP01, as it is the fundamental mode, SLM is not required to generate it. Fundamental mode LP01 and the excited higher order modes are shown in Figure 4. The utilization of these four modes is mentioned in table 1 below.

TABLE I
MODES UTILIZATION

Mode	Utilization
LP01	Downlink of RRU1
LP11a	Downlink of RRU2
LP21a	Uplink of RRU1
LP31a	Uplink of RRU2

Utilizing the MUX/DEMUX component in the simulator, four modes have been selected as the data channels. LP01 (Fundamental mode) and LP11a have been selected for the downlinks of RRU1 and RRU2 respectively. On the other hand, LP21a and LP31a have been selected for the uplinks. It was considered that each mode belongs to a different mode group to reduce the impact of the mode coupling that occurs during the propagation of these modes inside the FMF, Figure 4 shows the selected 4 modes and their initial powers.

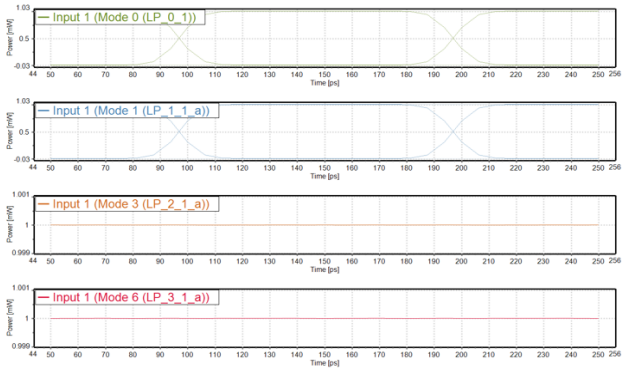


Fig. 4. Excited modes before launching them into the FMF at the CO side, LP01 and LP11a are modulated (Downlinks), LP21a and LP31a are carriers only (Uplinks)

D. Multiplexing Technique

The proposed setup considers Mode Division Multiplexing (MDM) as the multiplexing technique for the downlinks, where one mode carries the data towards its destination Remote Radio Unit (RRU). On the other hand, due to the bidirectional setup, the uplink modes suffer from a second round of mode coupling—a physical phenomenon that occurs during the propagation of light modes inside the fiber [1] — on their journey back to the Central Office (CO). To mitigate this impact, Mode Group Division Multiplexing (MGDM) has been selected as the multiplexing technique for the uplinks. Here, all coupled modes that belong to the same group are considered as a single data stream similar to the operating principle of the ordinary Multimode transmission systems.

IV. RESULTS AND DISCUSSION

A. Crosstalk

While the use of FMF reduces crosstalk between modes [1], it remains a factor during light mode propagation along the transmission fiber. Crucially, no crosstalk is observed with the fundamental mode LP01, as it belongs to a distinct mode group. However, our simulation setup reveals crosstalk between the degenerate modes LP11a and LP11b, and as shown in Figures 5 and 6.

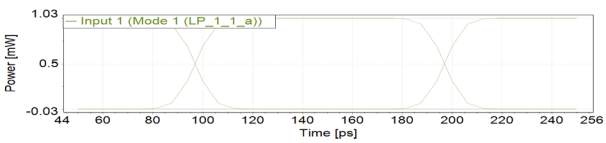


Fig. 5. Launched LP11a

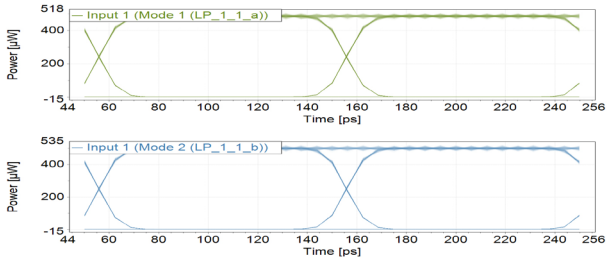


Fig. 6. Crosstalk between LP11a and LP11b

Moreover, severe crosstalk has been figured out in LP21a where its energy has been transferred to LP21b and LP02 and as shown in Figures 7 and 8.

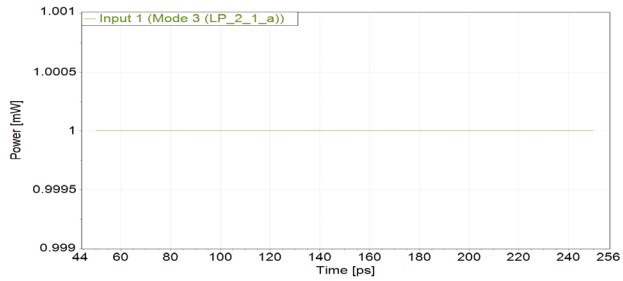


Fig. 7. Launched LP21a

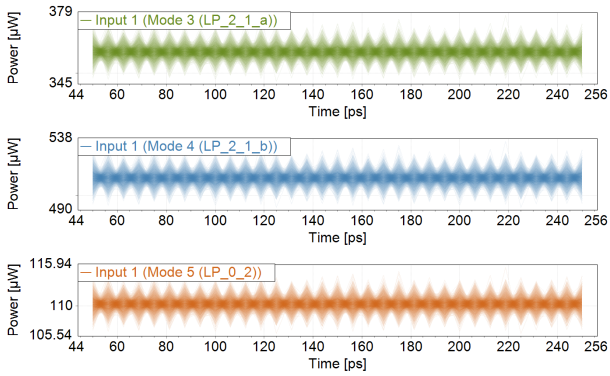


Fig. 8. Crosstalk between LP21a, LP21b, and LP02

For LP31a, higher energy transfer has been observed, energy has been transferred to LP31b, LP12a, and LP12b with more aggressive mode coupling impact and as shown in Figures 9 and 10.

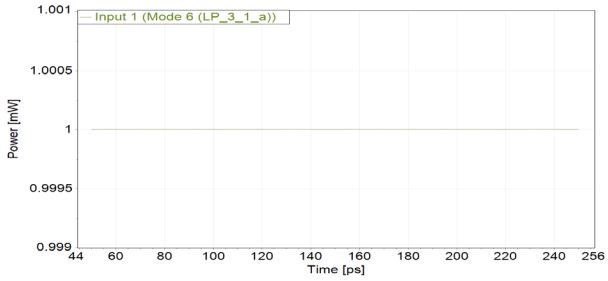


Fig. 9. Launched LP31a

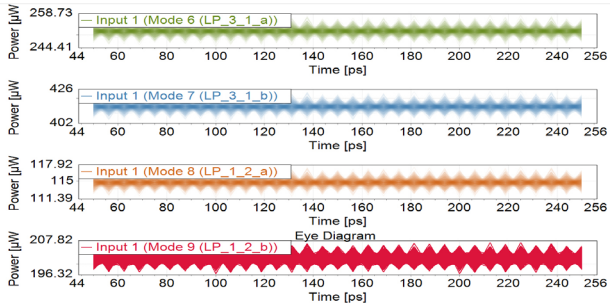


Fig. 10. Crosstalk between LP31a, LP31b, LP12a and LP12b

B. SLM Functionality

As mentioned in the previous sections, SLMs have been used for mode selection at the RRU side as a way to combat the crosstalk and ensure that only the mode of choice reaches its target destination, all other modes except the mode of choice are "filtered out", even if they belong to a different mode group. Figure 11 below shows an example of the SLM functionality to filter the unneeded mode LP11b.

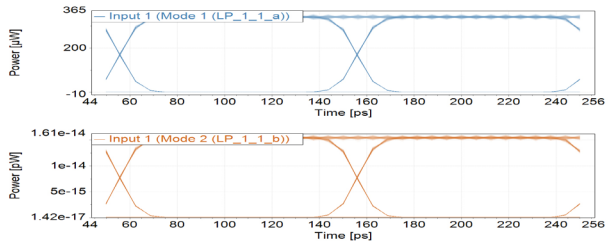


Fig. 11. SLM Functionality on LP11b

C. Analysis of the Eye Diagrams

Checking the obtained eye diagrams for the downlinks and the uplinks at 400m of fiber length, it is possible to say that adapting the concept of MDMA as a multiple access technique is feasible, the figures below show the obtained eye diagrams for the various links.

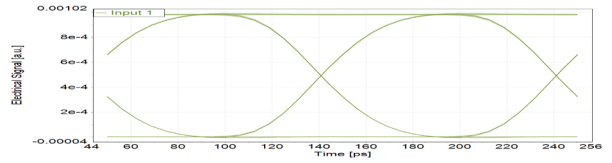


Fig. 12. RRU1 Downlink

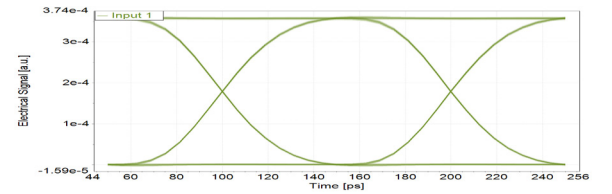


Fig. 13. RRU2 Downlink

The eye diagram openness indicates successful detection of the downlinks at the RRU side where high Optical Signal to Noise Ratio (OSNR) and low jitter are expected. Moreover, the implementation of MGDM shows promising results which can be expected with the obtained eye diagrams shown in the below figures.

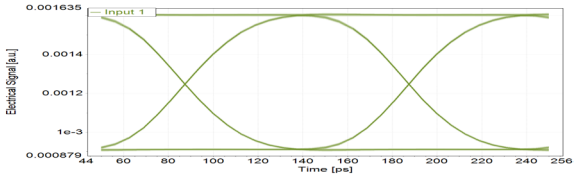


Fig. 14. RRU1 Uplink

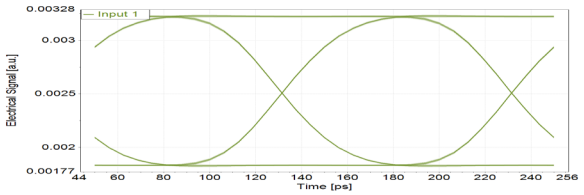


Fig. 15. RRU2 Uplink

V. CONCLUSION

In this study, we presented a proof of concept for using Mode Division Multiple Access (MDMA) as a viable alternative to the multiple access techniques available on the market. The modes were effectively generated and selected by the SLMs, after which they were multiplexed and demultiplexed on the RRU side. Furthermore, our proposed system can streamline RRU operations by centralizing the key mode generation process at the CO equipment. This paper shows the MDMA proof of concept through the simulation and analysis results and draws comparisons with our earlier works [1], [17], where the conceptual model and the behavior of the MDM system are elaborated. Additionally, we indicate that future research will explore reducing RRU complexities further by implementing the Power over Fiber (PWoF) approach.

VI. ACKNOWLEDGEMENT

The work has been supported by the National Research Development and Innovation Office (NKFIH) through the OTKA Grant K 142845.

REFERENCES

[1] A. S. Mohamed and E. Udvary, "Mode Division Multiplexing Bidirectional Transmission System using Double Clad Fiber for Mobile In-Building Solutions," in *2023 46th International Conference on Telecommunications and Signal Processing (TSP)*, 2023, pp. 144–147. DOI: 10.1109/TSP59544.2023.10197803.

[2] K. Zou, K. Pang, H. Song, J. Fan, Z. Zhao, H. Song, R. Zhang, H. Zhou, A. Minoofar, C. Liu, X. Su, N. Hu, A. McClung, M. Torfeh, A. Arbabi, M. Tur, and A. E. Willner, "High-capacity free-space optical communications using wavelength- and mode-division-multiplexing in the mid-infrared region," *Nat Commun*, vol. 13, no. 1, p. 7662, Dec. 2022. DOI: 10.1038/s41467-022-35327-w.

[3] J. Zhou, J. Wu, and Q. Hu, "Tunable arbitrary unitary transformer based on multiple sections of multicore fibers with phase control," *Opt Express*, vol. 26, no. 3, p. 3020, Feb. 2018. DOI: 10.1364/oe.26.003020.

[4] R. Mercy Kingsta and R. Shantha Selvakumari, "A review on coupled and uncoupled multicore fibers for future ultra-high capacity optical communication," *Optik (Stuttg)*, vol. 199, Dec. 2019. DOI: 10.1016/j.ijleo.2019.163341.

[5] P. Zhang, X. Xu, C. Dong, K. Niu, H. Liang, Z. Liang, X. Qin, M. Sun, H. Chen, N. Ma, W. Xu, G. Wang, and X. Tao, "Model division multiple access for semantic communications," *Frontiers of Information Technology & Electronic Engineering*, vol. 24, no. 6, pp. 801–812, Jun. 2023. DOI: 10.1631/FITEE.2300196.

[6] L. Wang, F. Jiang, Z. Yuan, J. Yang, G. Gui, and H. Sari, "Mode division multiple access: a new scheme based on orbital angular momentum in millimetre wave communications for fifth generation," *IET Communications*, vol. 12, no. 12, pp. 1416–1421, Jul. 2018. **DOI**: 10.1049/iet-com.2017.1182.

[7] L. Allen, M. W. Beijersbergen, R. J. C. Spreeuw, and J. P. Woerdman, "Orbital angular momentum of light and the transformation of Laguerre-Gaussian laser modes," *Phys Rev A (Coll Park)*, vol. 45, no. 11, p. 8185, 1992.

[8] M. Matsuura, A. Ikukawa, H. Kuboki, and Y. Kawamura, "Investigation of Crosstalk between Data Signals and Feed Light in Power-over-Fiber," *Photonics*, vol. 9, no. 6, p. 369, May 2022. **DOI**: 10.3390/photonics9060369.

[9] J. Yu, F. Tan, and C. Yu, "Few-Mode Fiber Characterization System Based on the Spatially and Spectrally Imaging Technique," *Sensors*, vol. 22, no. 5, p. 1809, Feb. 2022. **DOI**: 10.3390/s22051809.

[10] L. Wang, F. Jiang, Z. Yuan, J. Yang, G. Gui, and H. Sari, "Mode division multiple access: A new scheme based on orbital angular momentum in millimetre wave communications for fifth generation," *IET Communications*, vol. 12, no. 12, pp. 1416–1421, Jul. 2018. **DOI**: 10.1049/iet-com.2017.1182.

[11] R. Mercy Kingsta and R. Shantha Selvakumari, "A review on coupled and uncoupled multicore fibers for future ultra-high capacity optical communication," *Optik (Stuttg)*, vol. 199, Dec. 2019. **DOI**: 10.1016/j.ijleo.2019.163341.

[12] Y. Su, Y. He, H. Chen, X. Li, and G. Li, "Perspective on mode-division multiplexing," *Applied Physics Letters*, vol. 118, no. 20. American Institute of Physics Inc., 17-May-2021. **DOI**: 10.1063/5.0046071.

[13] V. Prajzler and M. Zikmund, "Power over fiber using a multimode optical power with a core diameter of 50 μm," *Opt Quantum Electron,* vol. 56, no. 8, p. 1307, Jul. 2024. **DOI**: 10.1007/s11082-024-07231-8.

[14] A. S. Mohamed and E. Udvary, "Enabling Enhanced In- Building Solutions Fronthaul Connectivity: The Role of Mode Division Multiple Access in Mobile Networks," in *2024 24th International Conference on Transparent Optical Networks (ICTON)*, 2024, pp. 1–5. **DOI**: 10.1109/ICTON62926.2024.10647457.

[15] K. Igarashi, D. Souma, T. Tsuritani, and I. Morita, "Performance evaluation of selective mode conversion based on phase plates for a 10-mode fiber," *Opt Express,* vol. 22, no. 17, p. 20881, Aug. 2014. **DOI**: 10.1364/oe.22.020881.

[16] F. A. Shnain and A. R. Salih, "Design and Study of Few- Mode Fibers at 1550 nm," *Journal of Educational and Scientific Studies*, vol. 18, no. 1, pp. 83–95, 2021.

[17] A. S. Mohamed and E. Udvary, "Modes selection in fronthaul Mode Division Multiplexing bidirectional transmission system for mobile In-building solutions," in *2nd Workshop on Intelligent Infocommunication Networks, Systems and Services*, Online, 2024, pp. 101–106. **DOI**: 10.3311/WINS2024-018.

[18] J. Carpenter, B. C. Thomsen, and T. D. Wilkinson, "Degenerate Mode-Group Division Multiplexing," *Journal of Lightwave Technology*, vol. 30, no. 24, pp. 3946–3952, Dec. 2012. **DOI**: 10.1109/JLT.2012.2206562.

[19] N. K. Fontaine, S. G. Leon-Saval, R. Ryf, J. R. Salazar Gil, B. Ercan, and J. Bland-Hawthorn, "Mode-Selective Dissimilar Fiber Photonic-Lantern Spatial Multiplexers for Few-Mode Fiber."

[20] J. von Hoyningen-Huene, R. Ryf, and P. Winzer, "LCoS- based mode shaper for few-mode fiber," *Opt Express,* vol. 21, no. 15, p. 18 097, Jul. 2013. **DOI**: 10.1364/OE.21.018097.

[21] W. Q. Thornburg, B. J. Corrado, and X. D. Zhu, "Selective launching of higher-order modes into an optical fiber with an optical phase shifter," *Opt Lett*, vol. 19, no. 7, p. 454, Apr. 1994. **DOI**: 10.1364/OL.19.000454.

[22] Z. Xuan, "A Research about Mode Multiplexing/Demultiplexing Technology," in *Proceedings of 2019 2nd World Symposium on Communication Engineering, WSCE* 2019, 2019, pp. 38–41. **DOI**: 10.1109/WSCE49000.2019.9040949.

Ahmed S. Mohamed received his B.Sc. from the university of Technology/ Iraq in 2007 and M.Sc. from the Universiti Utara Malaysia (UUM) in 2016. He is currently pursuing his PhD in Budapest University of Technology and Economics/Hungary in optical networks for mobile networks applications.

Eszter Udvary received her Ph.D. degree in electrical engineering from the Budapest University of Technology and Economics (BME), Budapest, Hungary, in 2009. She is currently an Associate Professor at BME, Mobile Communication and Quantum Technologies Laboratory. Dr. Udvary's research interests are in the broad area of optical communications, including microwave photonics, optical access network, visible light communication, and quantum communication.

On the Enhancement Anomaly Detection for RF Bio-Sensors by Computing Artificial Networks Using Machine Learning Techniques

Raya Adel Kamil[1], Saif Mohamed Baraa Alsabti[2], Rusul K. Abdulsattar[3], Ammar H. Mohammed[4], and Taha A. Elwi[5,*]

***Abstract*—This work proposes serval sensor designs for a low-cost, highly-sensitive microwave sensor for identifying different liquid samples by monitoring the variation in S21 magnitude. The sensor is developed using an interdigital capacitor (IDC) in series connection with a circular spiral inductor (CSI) and connected directly to a photo-resistor (LDR). To enhance sensor insertion losses, the sensor is introduced to a Hilbert fractal open stub and coupled to an interdigital capacitor to operate at 1.22GHz. The accuracy of the sensor is significantly improved using a back loop trace, eliminating nonlinear effects from multi-layer diffractions. An analytical model based on circuit theory is suggested for the proposed sensor operation. The authors found an observable influence of varying the LDR value on sensor insertion losses, motivating the development of the sensor prototype. The sensor is manufactured and tested experimentally before and after samples introduction, with a human glucose sample mounted on the LDR patch to measure the effects of light intensity.**

***Index Terms*—Microwave sensor, nondestructive biomedical measurements, Hilbert fractal structure, circular spiral, glucose, neural network.**

I. INTRODUCTION

Microwave sensors play a crucial role in biomedical applications, including noninversion sensing processes and analysis of bio tissue dielectric properties [1]. The future outlook for medicine is directed towards personal treatment regimens, which aim to establish personalized treatment plans for each patient [2]. Microwave technologies offer low-cost and low-power sensitivity, especially in the complex combination of compounds found in human body fluids like blood, glucose, and spinal fluids [3]. This leads to a wide time gap between sample acquisition and associated results. The increasing prevalence of chronic diseases and the need for cost-effective healthcare are the main challenges facing researchers in this field [4]. To achieve good medical care at low costs, researchers are focusing on spreading awareness of prevention and effective treatment against diseases rather than focusing on advanced treatment systems [5]. Healthcare providers and employers are increasingly adopting modern communication technologies, such as microwave sensors, to promote "e-health" monitoring, which provides accurate and early diagnosis without the need for external medical control [6].

Biosensors for microwaves depend on the characteristics that determine the electromagnetic fields interacting with materials based on their molecular structure [7]. Microwave biosensors are designed to mutate changes in wave spread speed during the biological environment into a quantifiable signal, providing the diversity of a specified bio-parameter [8].

RF/microwave resonators are essential in radio frequency (RF)/microwave frameworks for detection and quantization. For optimal channel plow insertion loss (IL), high return loss (RL), high frequency selectivity, low losses, more fetched, and compactness [6]. Microwave sensors advanced work with new sensing technologies to do multi-band operations. Such technology will make it possible for short-range, high-information-rate connections [8]. Microchannel planning commonly uses microwave resonators based on microstrip lines because of their cost-effectiveness, simplicity, and ease of manufacture [5]. To make multi-band microwave resonators, designers have used several different techniques, including stepping-impedance microstrip resonators, multi-mode resonators, parallel-coupled line resonators, and transmission zeroes [9]. To make sensors simpler and smaller, we can use stepped-impedance resonators, which work well with two or more transmission lines that have different characteristic impedances [10].

Biosensors are essential components in medical and biological experiments and diagnostics, measuring the dose of various biochemical species in aqueous solutions [11]. They have become essential in fields such as diagnostics, pharmaceutical procedures, biomedical engineering, industry, agricultural, and food safety [12]. Researchers have proposed various techniques and results for detection in different materials. In [13], a waveguide cavity-based sensor was presented for measuring the concentration of liquid solutions. The sensor operates at 1.91GHz in the fundamental TE101 resonant mode and has been tested on water-sodium chloride and water sucrose combinations. However, the accuracy of the resonant frequency evaluation is directly related to the quality factor QF and conditions around the measured records. In [14], a substrate integrated wave-guide (SIW) sensor was proposed to measure the permittivity of liquids. The resonant parameters of the sensor are affected by liquids passing through a slot

[1] Public Relations Department, College of media, Aliraqia University, Baghdad, Iraq, (e-mail: raadiff@gmail.com)
[2] Technical College of Engineering, Al-Bayan University, Baghdad, Iraq, (e-mail: saifsabbit@gmail.com)
[3] Electrical Engineering Department, University of Technology, Baghdad, Iraq, (e-mail: rusalhhf@gmail.com)
[4] Advisory Office for Scientific, Academic Affairs and Artificial Intelligence Applications, Iraqi Prime Minister Office, Baghdad, Iraq, (e-mail: ammmerhhaj@gmail.com)
[5, *] Department of Automation and Artificial Intelligence Engineering, College of Information Engineering at Al-Nahrain University, Baghdad, Iraq, (e-mail: taelwi82@gmail.com), the corresponding author is Prof. Dr. Taha A. Elwi

DOI: 10.36244/ICJ.2025.2.11

opened on the upper side of the sensor. The sensitivity sensor is tested on different percentages of water in ethanol. The arterial neural network was used to solve the problem with relative errors of 5% and 7%, respectively. In [15], a microwave resonator based on Complementary Circular Spiral Resonator (CCSR) was proposed, working at 4.72GHz. Through experimentation and analysis, the proposed sensor can determine the concentrations of ethanol-water mixtures by measuring the resonant frequency of the CCSR and the permittivity of sample under test (SUT). In [16], a microwave resonator based on the complementary split-ring resonator (CSRR) coupling with a microstrip and the microfluidic channel was proposed, working at 3.994GHz. The sensor can also determine the water content of glucose by measuring the resonant frequency of the CSRR and the permittivity of SUT. In [17], a microwave sensor with a coplanar waveguide semi-lumped meandering open complementary split ring resonator (MOCSRR) was proposed, operating up to frequencies of approximately 200MHz. The sensor successfully detects branded and unbranded fuel oil samples, with the difference demonstrated by the fluctuation in the transmission coefficient resonant frequency amplitude. In [18], a Cesare fractal geometry based on a compact Electromagnetic Bandgap (EBG) structure was presented to measure the complex permittivity of various liquids (butan-1-ol, methanol, and water). The relative permittivity is 3.57 for butan-1-ol, 21.3 for methanol, and 78 for water.

In this work, the proposed sensor is realized for enhancing anomaly detection of RF bio-sensors using machine learning techniques. For this the theoretical considerations and design are discussed in section II. The experimental validation is considered in section III. In section IV, the neural network implementation is realized. The paper is concluded in section V.

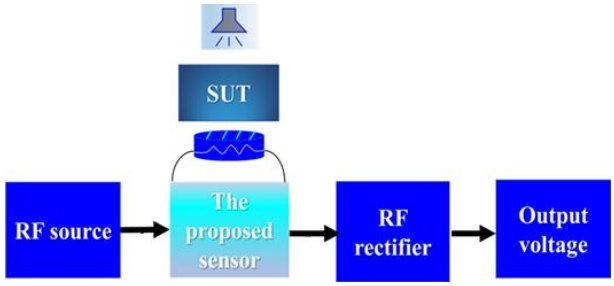


Fig.1: The simplified block diagram of the design based on reconfigurable technology.

II. THEORETICAL STUDY TO DESIGN A MICROWAVE SENSOR FOR BIOMEDICAL DETECTIONS

This paper presents new microwave sensors for biomedical detection using a two-port network for liquid characterizations. The main structure is a miniaturized microwave resonator based on a circular spiral inductor, which is used to increase sensitivity and concentrate current before being transferred to the SUT. The sensor is introduced to three inclusions: open stub, Minkowski filter, and Hilbert filter. Open stubs are used to eliminate measurement distortions and dispersion effects, while Minkowski filter is used to linearize measurements and remove the effects of frequencies before 0.5GHz. The Hilbert curve is used to remove the effects of frequencies before 0.5GHz. The sensors are designed to be compatible with microprocessors and measure water content variation based on frequency shift. The research also explores the use of a photo resistor to control sensor performance and measure output voltage from a RF rectifier at the sensor's output port. A neural network model is presented to solve problems involving nonlinearities, multi-variables, and multi-resonances.

A. Base Sensor Design

The proposed sensor is mounted on a FR4 substrate with a thickness of 1.6 mm and is based on OS-CRLH. It consists of a transmission line connected to an RLC branch network, which is structured as input capacitor which are inspired from [19] in series with a circular spiral inductor Lse and connected directly to a photo-resistor LDR. The LDR is mounted between the LC branches at the middle of the sensor, and the back panel is covered with a metallic ground plane of 0.035 mm. The proposed sensor geometrical details are explained as seen in Fig.2. Consequently, the main structure of the proposed sensor is constructed from the same proposed sensor with Hilbert curve introduction. As maintained later, the proposed sensor is constructed from an inter digital capacitor to remove the effects of the imaginary component that is generated by the indictor structure which mainly stores the energy from the propagation [19]. The inductor structure is invoked to be a spiral geometry cause of being a highly sensitive stricter that concentrate the current before being transferred to SUT.

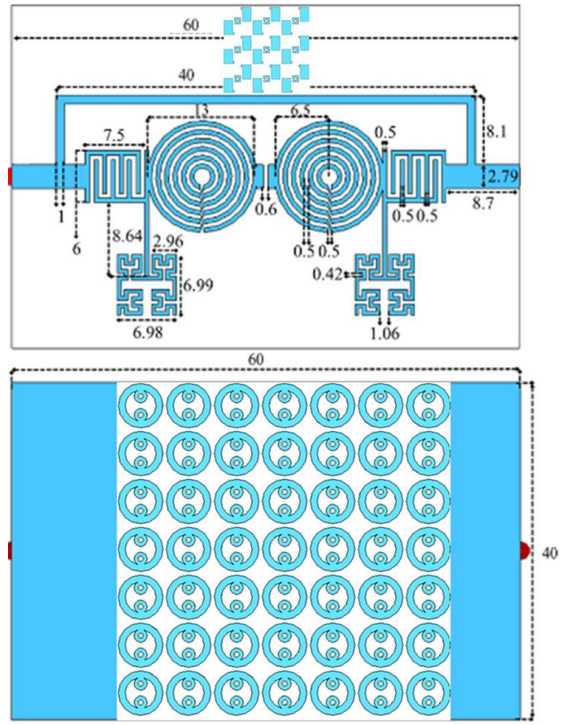


Fig.2; The proposed sensor geometry: (a) front view and (b) back view. Note: all dimensions are in mm scale.

The suggested design examines the Hilbert-curve configuration to minimize sensor dimensions and enhance frequency resonances [20]. The examined fractal geometry is founded on the third-order Hilbert geometry, as seen in Fig. 2.

The suggested equivalent circuit model for Hilbert geometry, as shown in Fig. 2, is developed from Richard Koch's theory [8]. The equivalent circuit model of the proposed unit cell is produced from the simulation results, as seen in Fig. 2. The equivalent capacitances of coupling between the unit cell and neighboring cells are denoted as the left-hand capacitor (CST), while the fractal slot is regarded as an inductor (LST), where the magnetic current mobility in the air traces may be amplified. This inductor corresponds to the magnetic field contained inside the fractal slots of the rings. The material loss is determined by the resistor RST [11]. The suggested sensor design has a feedback loop structure. The benefit of this introduction is to get a band reject filter response rather than a passband resonance. This alteration enhances the accuracy of the measurement findings [12].

The suggested sensor design is founded on an equivalent circuit model, which is analytically derived from an analogous circuit model based on the established RLC network, typically configured as an IDC in series with a CSI and the Hilbert fractal. The suggested structural equivalent circuit model is developed with the lumped elements Richard model [1]. The suggested circuit model consists of a 50Ω input impedance RF source connected in series with a parallel (R-L-C) branch, as seen in Fig.3. The primary transmission line was characterized by an inductive segment LT and capacitive air gaps Cgap, as previously seen in Fig.3(a). The proposed circuit model's S-parameters, shown in Fig. 3(b), are analyzed and juxtaposed with those obtained via CST MWS. An effective agreement is attained based on the specified lumped components, which are modeled in Advanced Devices Simulator (ADS). The assessed RLC components are enumerated in Table I.

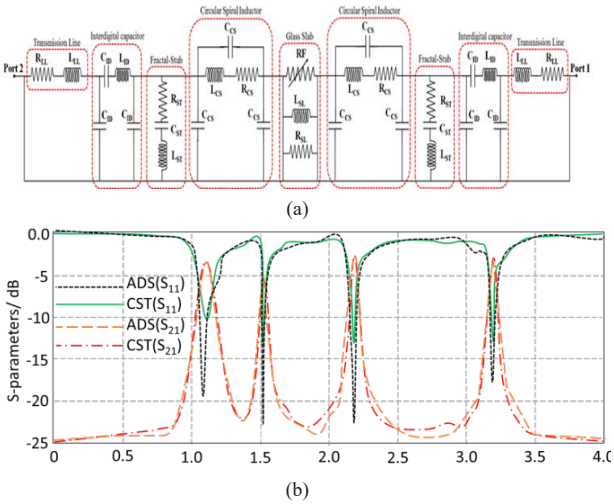


Fig.3 Equivalent circuit analytical model of the proposed sensor: (a) circuit model and (b) S-parameters.

TABLE I
LUMPED ELEMENT VALUES OF THE EQUIVALENT CIRCUIT MODEL

Element	RL H	RR H	GL H	GR H	CL H	CR H	LL H	LRH
Valus	12.2 Ω	50 Ω	0.1 S	4 S	1.1p F	3.1p F	3nH	2.2n H

C. Sensor Operation and Detection Process

This section presents the operational and reconfiguration scenarios designed to illustrate the proposed technique for comprehending sensor functionality. By altering the resistance of the included LDR from 100Ω to 1000Ω, a notable shift in the spectra of the suggested sensor S21 is seen, as shown in Fig. 4. This change is ascribed to the voltage division effects between the overall impedance of the proposed sensor and the LDR [13]. The present motion would be substantially influenced and traverse the back loop structure to be diffused inside the suggested fractal form. Consequently, this dissipation will be very beneficial for sensing, as will be shown subsequently. Fig.4 illustrates three frequency resonances at 630MHz, 1.22GHz, and 3.2GHz, all of which will be used in the sensing procedure of this study.

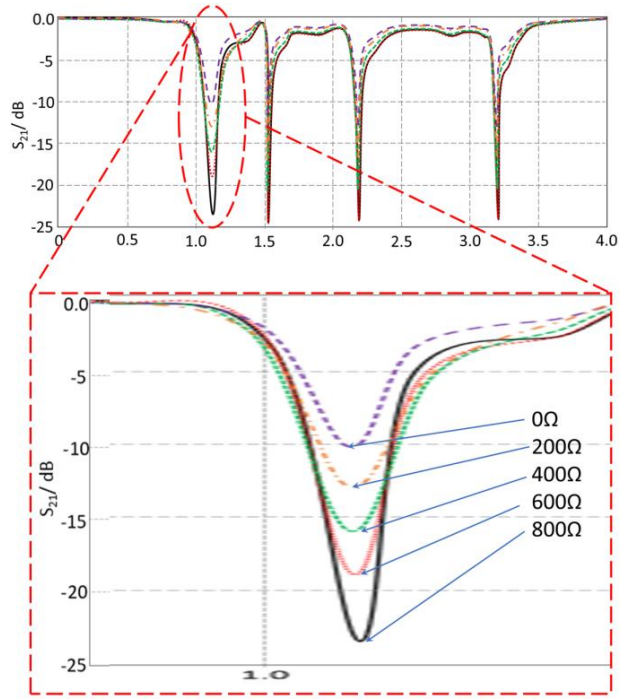


Fig.4: Evaluated S_{21} spectra with respect to varying the LDR value.

III. EXPERIMENTAL VALIDATIONS

This paper presents a sensor design for glucose samples aimed at improving detection accuracy by generating an output voltage. The main limitation is the difficulty in linearizing measurements due to field fringing from boundary conductions. The proposed method uses a photo-resistor to control sensor performance, eliminating field fringing effects from discontinuities. The sensor is connected with a back loop structure, allowing for band reject filter response and controlling charging rise time until the photo-resistor reaches saturation. The sensor is fabricated and measured experimentally using a Professional Network Analyzer (Agilent PNA 8720) after a through transmission calibration process. The measurements are conducted to S11 and S21 spectra to eliminate possible errors. The sensor shows a well-defined resonance at 1.5GHz with S21= -27 dB. An excellent agreement is found between numerical results and measurements within frequencies from 0.1GHz up to 4GHz.

The sensor was tested by taking different samples of the glucose and water mixture and measured experimentally in terms of S21 spectra. The use of an LDR switch is conducted with a long glass slab to avoid direct contamination and linearize variation using the LDR switch. The sensor is fabricated using a wit chemical etching process and S21 spectra are measured before and after glucose introduction on top of the LDR spot using an Agilent vector network analyzer. For this, an experimental study was conducted to validate the effectiveness of glucose level variation on sensor performance at 1.22GHz. The results showed that the proposed sensor is an excellent candidate for glucose measurements and could be promising for other biological fluid characterizations. The sensor was introduced to 15 patients and measured glucose levels at three different times, about 7 days to 10 days apart. The results showed that the variation in output voltage increased rapidly with increasing glucose magnitude as listed in Table II. The glucose sample used was about 0.01cc to avoid contamination effects.

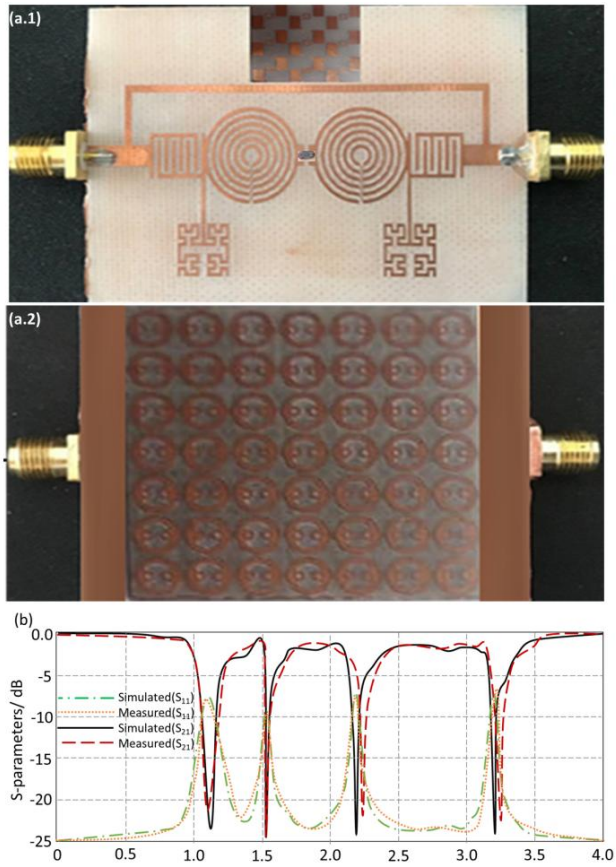
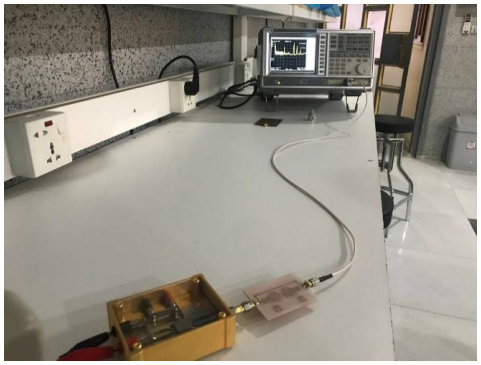


Fig.5; Experimental validation: (a) Fabricated sensor and (b) S-parameter spectra.

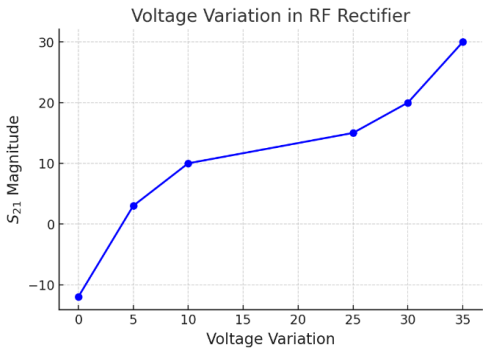
The proposed sensor is linked to an RF rectifier to gauge the output voltage across a resistor connected in parallel to the output terminals of the RF rectifier, as shown in Fig. 6(a). This measurement is conducted by varying the light intensity on the photoresist from 100 Ω to 600 Ω. This variance is analyzed quantitatively by examining the effective impedance change and its impact on the S21 value at 1.22GHz. The fluctuation is experimentally monitored about an output voltage (Vout) for practical use. The fluctuation in Vout has a linear trend, as seen in Table II. This signifies that a satisfactory concordance has been attained between the experimental and simulated outcomes. The proposed sensor variation in response may be attained by using an RF rectifier to assess the alteration in output voltage via an oscilloscope. The input voltage is set at 100 mV from the sources. Fig.6(b) illustrates the fluctuation of the output voltage of the proposed RF rectifier in relation to the change in input power.

TABLE II
MEASURED GLUCOSE INFLUENCE ON THE PROPOSED SENSOR PERFORMANCE.

Case number	BMI	Age/ year	Sex	Glucose level	mV	|S21|
1	19.1	7	F	102	72.2	0.106
				200	74.7	0.115
				156	73.9	0.111
2	24.9	51	F	111	72.8	0.109
				123	72.3	0.112
				201	75.3	0.115
3	21.4	70	M	300	77.2	0.121
				125	72.9	0.111
				245	76.3	0.156
4	28.1	45	M	301	76.9	0.116
				359	78.1	0.182
				277	78.4	0.174
5	26.7	56	M	231	74.5	0.123
				93	72.1	0.193
				108	72.4	0.191
6	32.1	35	F	122	73.1	0.103
				133	73.1	0.109
				143	73.9	0.108
7	33.5	43	F	185	74.1	0.122
				164	73.9	0.133
				166	72.4	0.124
8	29.8	42	F	101	73.1	0.091
				91	72.9	0.094
				98	73	0.092
9	23.4	49	F	144	74.1	0.091
				187	77.8	0.098
				145	77.4	0.092
10	32.6	37	M	190	76.3	0.109
				123	75.6	0.106
				144	76.1	0.11
11	36.1	72	M	102	72.4	0.111
				300	77.3	0.11
				340	77.9	0.113
12	34.6	67	M	390	77.8	0.189
				331	77.3	0.188
				301	77	0.177
13	21.4	63	M	210	74.9	0.109
				243	75.2	0.101
				226	75.1	0.105
14	22.5	68	M	189	74.2	0.195
				130	73.4	0.196
				210	74.4	0.179
15	23.8	54	M	221	75.3	0.109
				289	75.9	0.11
				234	75.2	0.112



(a)



(b)

Fig.6: (a) Experimental setup and (b) RF rectifier performance.

IV. NEURAL NETWORK IMPLEMENTATION

In this section, the variation in the measured S_{21} magnitude from is performed. The S_{21} variation with respect to the glucose level normalization are evaluated from measurement data. In such data, the detection process is performed according to the S_{21} change. For this, a comparison study between measured data is conducted to realize the trade-line regressions for measurements with a negative slop as seen in Fig.7.

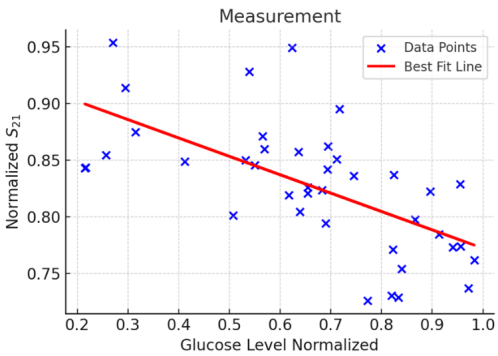


Fig.7: A comparison study between measured points.

This study used the instances from Table II to derive the input data. The samples are categorized into training and testing sets. To find the best performance index (P.I.) for the ANN, the number of neurons, epochs, and learning rate will be changed. The neural network will then be trained five times, and the results will be averaged. There is one buried layer containing three neurons. The learning rate is set at 0.001, with a total of 88 epochs. The mean P.I. of this network is 76.32%. The MATLAB code is used to categorize the input data from regression, demonstrating the categorization of the data based on their respective categories. The input data in this category is classified by the neural network into three periods, mostly based on the regression rate. In this categorization, the first third of the input data is designated as low glucose level. The second interval pertains to the intermediate glucose concentration. The last interval is regarded for elevated glucose levels. Figure 8 demonstrates that the data regression aligns very well with the output data. Furthermore, the regression topic is notably relevant; the suggested sensor, based on the neural network, achieves exceptional alignment with classifications at both low and high glucose levels. This finding is confined to intermediate values, which may result in significant inaccuracy. To provide an effective solution, more data points are necessary to identify the optimal fit for this period. It is noteworthy that the disparity in the intermediate period is shown in Fig. 7, which depicts a breakpoint in the center of the values from the simulation that aligns with the actual data. Table III enumerates the optimal values achieved for the most frequently used neural network parameters.

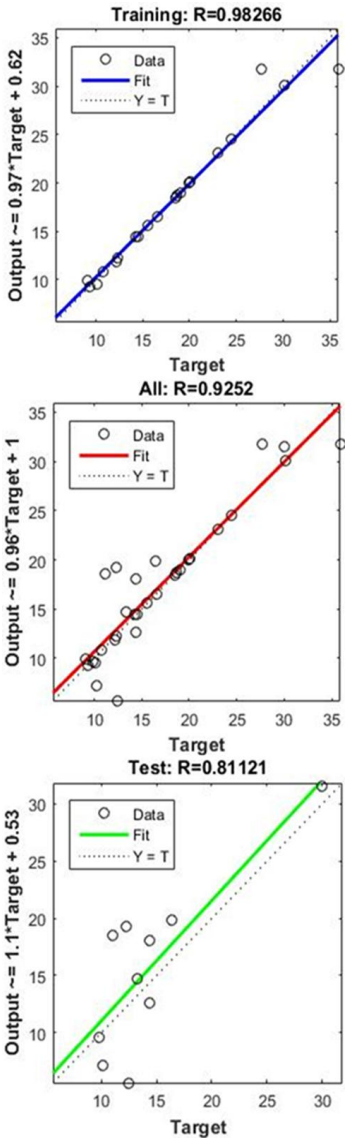


Fig.8: Findings from the regression analysis.

TABLE III
NN PARAMETERS

NN parameters	NN for simulation
Number of Input layer nodes	5
Number of Hidden layer nodes	3
Number of Output Layer nodes	1
Transfer function	logsig
Training function	pureline
Learning rate	0.001
Maximum number of Epochs	88

Finally, the proposed sensor is compared to the relatives published in the literature as listed in Table IV. The proposed work is compared interms of Q-factor, sensing type, and frequency. It is found that the proposed sensor provides the highest Q-factor with a resonance frequency of 1.22GHz. This frequency band is highly suitable for biomedical applications, particularly glucose level detection. Because it can balance penetration depth, sensitivity, and signal integrity, the 1.22GHz frequency is perfect for biomedical applications that need to find glucose levels [12]. Its dielectric properties, which vary with frequency, make it easier to detect variations in glucose concentrations. The frequency also keeps signal loss to a minimum, which makes it possible to use a non-invasive method that is still sensitive to changes in glucose levels. Additionally, it offers biomedical safety and non-invasiveness, as it doesn't require direct contact with blood. The high-Q sensor at 1.22GHz allows for compact microwave sensors with strong resonance characteristics, improving measurement precision.

TABLE IV
A COMPARISON BETWEEN THE PROPOSED SENSOR AND OTHER PUBLISHED RESULTS

Ref.	Q-factor	Sensing	Fo/GHz
[4]	280	Solid	2.4
[5]	407.34	Solid	2.2
[6]	345	Solid	3.2
[7]	652	Solid	2.22
[8]	446, 506	Solid	2.5 and 3.9
[9]	458	Solid	1.5
[10]	662	Solid	2.4
[11]	265	Liquid	2.45
[12]	398	Liquid	1.8, 2.45, and 3.5
[13]	425	Liquid	5.3 and 5.8
[14]	280, 160	Liquid	5.76 and 7.85
[15]	111.56, 21.39	Liquid	2.45 and 5.8
[16]	286.5	Powder	1 to 3
[17]	385.6	Powder	1.0–3
The proposed work	794.7	Liquid	1.22

V. CONCLUSION

The proposed sensor design utilizes a microwave resonator based on the CRLH structure of Hilbert geometry. It is tested with glucose from 11 different patients, and the way it works involves a new way of using an LDR part that makes the sensor work differently when the amount of glucose changes. We attribute this change to the transparency of the glucose under test, which alters the frequency shift and S21 magnitude. The writers found that the ratio in the S21 magnitude is critical at 1.22GHz with linear variation. Because of the linear variation, the authors believe that this design is ideal for sensing. The proposed sensor circuit model is used to see what happens when the proposed sensor parts are added, and the outcomes are contrasted with the actual outcomes that were measured. We numerically test the sensor's performance using CST MWS and validate it with ADS. It was solved analytically with circuit model analysis, and the suggested sensor works in a straight line, but the way it works changes depending on what is being tested. Transmission line technology forms the suggested sensor. It has a transmission line that is linked to an RLC network, set up as an IDC in series with a CSI, and linked directly to an LDR. The sensor operates at 1.22GHz and detects water introductions successfully by changing the S21 magnitude directly.

REFERENCES

[1] Arkan Mousa Majeed, Fatma Taher, Taha A. Elwi, Zaid A. Abdul Hassain, Sherif K. El-Diasty, Mohamed Fathy Abo Sree, & Sara Yehia Abdel Fatah. (2024). High Gain Defected Slots 3D Antenna Structure for Millimeter Applications. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 46(1), 136–145. DOI: 10.37934/araset.46.1.136145.

[2] Raghad Al-Shabandar, Ali Jaddoa, Taha A. Elwi, A. H. Mohammed, and Abir Jaafar Hussain, "A Systematic Review for the Implication of Generative AI in Higher Education", *Infocommunications Journal*, Vol. XVI, No 3, September 2024, pp. 31-42., DOI: 10.36244/ICJ.2024.3.3.

[3] R. K. Abdulsattar, T. A. Elwi and Z. A. A. Hassain, "A New Microwave Sensor Based on the Moore Fractal Structure to Detect Water Content in Crude Oil," *Sensors*, vol. 21, no. 21, p. 7143, 2021. DOI: 10.3390/s21217143.

[4] A. Al-Behadili, I. A. Mocanu, T. M. Petrescu and T. A. Elwi, "Differential Microstrip Sensor for Complex Permittivity Characterization of Organic Fluid Mixtures," *Sensors*, vol. 21, no. 23, p. 7865, 2021. DOI: 10.3390/s21237865.

[5] R. K. Abdulsattar, T. A. Elwi and Z. A. A. Hassain, "Microwave Resonator based a Fractal Moore Structure for Modern Wireless Reconfigurable Systems," *2021 3rd Int. Conf. on Research and Academic Community Services (ICRACOS)*, pp. 98–102. DOI: 10.1109/ICRACOS53680.2021.9701943.

[6] Yahiea Alnaiemy, Mohammed N. Majeed, Mohsin Ali Ahmed, Taha A. Elwi, and Sarah Mohemmed Fawzi Hussein, "Intelligent Antenna Array Systems for Modern Communication Networks", *Acad. Sci. J.*, vol. 3, no. 1, pp. 1–15, Jan. 2025, DOI: 10.24237/ASJ.03.01.946E.

[7] M. Mahdi Ali, E. Márquez Segura, and T. A. Elwi, "Advancements in Ku Band Resonator Composite Right/Left-Handed (CRLH) Metamaterials: Design, Analysis, and Applications", *J. eng. sustain. dev.*, vol. 29, no. 2, pp. 184–189, Mar. 2025, DOI: 10.31272/jeasd.2646.

[8] Fatma Taher, Wasan S. Rasheed, Taha A. Elwi, Hayder H. Al-khaylani, Mohamed Fathy Abo Sree, Sara Yehia Abdel Fatah, M. S. H. Salah El-Din. (2025). Novel Reconfigurable Fractal Antenna Design for Modern Communication Systems. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 64(4), 47–57. DOI: 10.37934/araset.64.4.4757.

[9] Ahmed F. Miligy, Fatma Taher, Taha A. Elwi, M. Abdelaleim, Mohamed Fathy Abo Sree, and Sara Yehia Abdel Fatah, "Broad Band 2×4 Horn Antenna Array for High Power Microwave (HPM) Systems Application", *J. Adv. Res. Appl. Sci. Eng. Tech.*, vol. 64, no. 4, pp. 1–17, Mar. 2025, DOI: 10.37934/araset.64.4.117.

[10] Fatma Taher, Adham R. Azeez, Taha A. Elwi, Mohamed Fathy Abo Sree, Ahmed F. Miligy, M. Abdel-aleim M., & Sara Yehia Abdel Fatah. (2025). Design of Offset Radiation Tapered Slot Antenna. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 64(4), 37–46. DOI: 10.37934/araset.64.4.3746.

[11] Raya Adel Kamil, Noof T. Mahmood, Zainab Salam Muqdad, Marwah Haleem Jwair, Noor Mohammed Noori, and Taha Ahmed Elwi, "On the Performance of Metasurface Vivaldi Antenna in Breast Cancer Detection Using Artificial Neural Networks for Bio-Signal Analysis," *Progress In Electromagnetics Research B*, Vol. 111, 31-43, 2025, DOI: 10.2528/PIERB24122803.

[12] Al-Gburi, Rasool M., Alibakhshikenari, Mohammad, Virdee, Bal S., Hameed, Teba M., Mariyanayagam, Dion, Fernando, Sandra, Lubangakene, Innocent, Tang, Yi, Khan, Salah Uddin and Elwi, Taha A.. "Microwave-based breast cancer detection using a high-gain Vivaldi antenna and metasurface neural network approach for medical diagnostics" *Frequenz*, 2025. DOI: 10.1515/freq-2024-0190.

[13] Taher, F., Bashar, B. S., Elwi, T. A., Misran, H., Rhazali, Z. A., Ismail, M. M., Abdel Fatah, S. Y. (2025). Novel antenna-based metamaterial structure with Slotted and parastic patches for 5G-sub 6 GHz applications. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 65(2), 143–152. DOI: 10.37934/araset.65.2.143152.

[14] S. Ifzarne, H. Tabbaa, I. Hafidi and N. Lamghari, "Anomaly Detection using Machine Learning Techniques in Wireless Sensor Networks," *Journal of Physics: Conference Series*, vol. 1743, no. 1, p. 012021, 2021. DOI: 10.1088/1742-6596/1743/1/012021.

[15] Hamad, A. R., Al-Adhami, A., Elmunim, N. A., Alibakhshikenari, M., Virdee, B., Hamad, H. S., et al. (2025). Rectenna design optimized by binary genetic algorithm for hybrid energy harvesting applications across 5G sub-6 GHz band. *Radio Science*, 60, e2024RS008154. DOI: 10.1029/2024RS008154.

[16] Rashid, S. A., Audah, L., Hamdi, M. M., Abood, M. S., Abbas, G. R., Mohammed, B. S., et al. (2025). Delay-minimization and back-off aware Q-learning with advanced bio-inspired CH selection for multi-hop communication in vehicular ad-hoc networks. *Radio Science*, 60, e2024RS008165. DOI: 10.1029/2024RS008165.

[17] D. Ground and J. Yeo, "High-Sensitivity Microwave Sensor Based on an Interdigital-Capacitor-Shaped Defected Ground Structure for Permittivity Characterization," *Sensors*, 2019, 19(3), 498. DOI: 10.3390/s19030498.

[18] R. A. Alahnomi, Z. Zakaria, Z. M. Yussof, T. Sutikno, H. Sariera and A. A. M. Bahar, "Accurate characterizations of material using microwave t-resonator for solid sensing applications," *Telkomnika*, vol. 18, no. 1, pp. 99–105, 2020. DOI: 10.12928/TELKOMNIKA.v18i1.14880.

[19] S. Y. Jun, B. Sanz Izquierdo and E. A. Parker, "Liquid Sensor/Detector Using an EBG Structure," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 5, pp. 3366–3373, 2019. DOI: 10.1109/TAP.2019.2902663.

[20] S. Kulkarni and M. S. Joshi, "Design and Analysis of Shielded Vertically Stacked Ring Resonator as Complex Permittivity Sensor for Petroleum Oils," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 8, pp. 2411–2417, 2015. DOI: 10.1109/TMTT.2015.2451110.

Raya Adil Kamil is a dedicated professional in the field of computer communication engineering and computer science, with a strong passion for advancing her skills and expertise in cutting-edge technologies. She holds a Bachelor's degree in Computer Communication Engineering from Al-Rafidain University College (2006-2007) and later earned her Master's degree in Computer Science from the College of Information Technology, Middle East University, Amman, Jordan (2016-2017). With a career spanning multiple roles in academia and the public sector, Raya has gained valuable experience in technology, information security, and higher education. She worked at the Iraqi Ministry of Planning, specifically in the Department of Technology and Information Security, from April 2010 to March 2012. She then joined the International Leaders Academy in Amman, Jordan, where she contributed to the Department of Information Technology from December 2014 to October 2014. Raya has also been active in academia, serving as a lecturer at various institutions. She was a lecturer in the Department of Information Technology at Ummah University, Istanbul, Turkey, from March 2017 to October 2017. More recently, she worked as a lecturer at Al-Turath University College in the Department of Medical Devices Technology Engineering for the 2022-2023 academic year. She holds the scientific title of Teacher Assistant. Currently, Raya is seeking opportunities in a large company where she can enhance her skills, gain experience with advanced technologies, and contribute to innovative projects. She is eager to bring her expertise in computer science, information security, and technology management to a dynamic and growth-oriented organization.

Saif Mohamed Baraa Al-Sabti was born in Iraq in 1980. He received two B.Sc. degrees: one in Biochemical Engineering from Al-Khwarizmi Engineering College, Baghdad University, in 2005, and another in Com-munications Engineering from Al Ma'mon University College, Baghdad, Iraq, in 2017. He earned his Ph.D. degree in Electrical and Computer Engineering from Altınbaş University, Turkey, in 2021. Dr. Al-Sabti has authored several published papers in the fields of sensors, communications, antennas, SARs, and electromagnetics. His research contributions have been influential in advancing these areas, and he has also taught at Al-Turath University College. Currently, he is a lecturer at the Medical Devices Techniques Engineering Department, College of Engineering Technology at Al-Bayan University, Baghdad, Iraq. Dr. Al-Sabti continues to inspire and mentor the next generation of engineers through his teaching and research efforts.

Rusul K. Abdulsattar received her B.Sc. degree in Electrical Engineering from College of Engineering at A Mustansiriyah University, Baghdad, Iraq, in (2017). She obtained her M.Sc. degree in electronic and communications engineering from Mustansiriyah University, Baghdad, Iraq in (2022). She worked as an assistant lecturer at Al-Hadi University College, Baghdad, Iraq in (2022). Currently, she's an Assist. Lect. in the University of Technology. Her research interests include microwave design, microstrip filters, reconfigurable technology, and artificial neural networks.

Ammar H. Mohammed received the B.S.C. in Physics Science from University of Mustansiriyah-Iraq and M.S.C in Physics Science – electron optics (2014) from University of Mustansiriyah (2004). Dr Ammar holds a PhD in Nano Optics and is affiliated with the Prime Minister's Office for Scientific Research Affairs and Al Applications, where he serves as the Rapporteur of the Supreme Committee on Al Applications. He has conducted extensive scientific research in nanotechnology related to Al applications. His work has been recognised with a research grant from the Russian Foundation for Basic Research and another from the Russian Science Foundation for nano-thermometry therapy applications. Dr Ammar has numerous certificates in crime scene investigation and forensic evidence analysis from the British Embassy, USA. He is also certified in the use of the AFIS system for fingerprint comparison by Getx. Additionally, he holds various certificates in chemical analysis, particularly in GC Mass systems, FTIR, and atomic absorption.

Taha A. Elwi is a distinguished researcher and academic leader in the fields of electromagnetics, wireless communications, and biomedical applications. He received his B.Sc. in Electrical Engineering from Al-Nahrain University, Baghdad, Iraq, in 2003, earning the Highest Graduation Award. He further pursued his M.Sc. in Laser and Optoelectronics Engineering at the same institution, graduating in 2005 with another Highest Graduation Award. From April 2005 to August 2007, Dr. Elwi worked with Huawei Technologies Company in Baghdad, gaining industry experience in telecommunications. In January 2008, he joined the University of Arkansas at Little Rock, where he earned his Ph.D. in System Engineering and Science in December 2011. His outstanding contributions to research earned him recognition as one of Stanford University's Top 2% Scientists in 2022. Dr. Elwi's research spans multiple advanced domains, including: Wearable and implantable antennas for biomedical wireless systems, Smart antennas and WiFi deployment, Electromagnetic wave scattering by complex objects, Metamaterial structures for microwave applications, Microstrip antennas for mobile radio systems, Precipitation effects on terrestrial and satellite communication systems, Electromagnetic propagation and GPS in complex media, Wireless sensor networks using microwave and laser optoelectronic devices, Nano-scale structures across the electromagnetic spectrum, Reconfigurable intelligent surfaces (RIS) and channel performance control, Cognitive radio networks and squint effects in modern wireless communications, Pattern recognition, signal and image processing, machine learning, deep learning, game theory, and AI-based medical image analysis. As a respected figure in academia and research, Dr. Elwi serves as an editor for prestigious publishers such as MDPI, IEEE, Springer, and Elsevier. He is also the Head of the International Applied and Theoretical Research Center (IATRC), Baghdad Quarter, Iraq, and has been a member of the Iraqi Scientific Research Consultant Committee since 2016. Dr. Elwi leads three international collaborations focused on biomedical applications using microwave technology and supervises numerous funded projects and Ph.D. theses. His prolific research output includes over 150 published papers and 10 patents, making him a highly influential expert in his field.

Constrained LS Channel Estimation for Massive MIMO Communication Systems

Muhammad Ahsan Shaikh, Tayyab Ahmed Shaikh, Sadiq ur Rehman, and Halar Mustafa

***Abstract*—In recent years, the manufacturing of mobile and IoT devices has increased dramatically. For the service provider, the requirement for high throughput and extensive connectivity became a major obstacle. In B5G and 6G, different advanced technologies have been introduced to cater demands of users effectively. One of the most important technologies of next-generation networks is massive MIMO systems. In multiuser communication systems, transmission and reception of signals occur simultaneously which creates multiuser interference (MUI). The presence of MUI in the system is the major challenge for the effective operation of massive MIMO receivers. The influence of MUI must be minimized using a channel estimation technique in order to fully utilize the capabilities of a massive MIMO system. This work proposes the constrained least square (LS) channel estimate technique to improve the massive MIMO downlink system's overall performance. The Mean Square Error shows that the unconstrained LS performance is poor as compared to the constrained LS channel estimation. Additionally, the effectiveness of the proposed constraint LS channel estimate is assessed in communication systems using varying transmission antennas at the base station and number of users.**

***Index Terms*—Massive MIMO, Channel Estimation, B5G, MUI, Least Square**

I. INTRODUCTION

In the era of globalization, contemporary networks encounter a significant upswing in traffic demand. To meet the specific demands of cellular systems, these networks are strategically deployed over short distances. Additionally, wireless Local Area Networks (WLANs) are ubiquitously employed in various locations to address connectivity needs. The growing popularity of mobile broadband services and the introduction of new concepts like machine-to-machine (M2M) and the Internet of Things (IoT) are also responsible for the rise in wireless traffic. In daily routine, consumers are more likely to rely severely on mobile data due to the widespread of cellular services. Our living standards have increased thanks to the greater potential that the 3G, 4G, and 5G have brought forth, such high data rates and minimal latency. The improved feature in new generations [1], has enabled users to do online gaming, video calls, and engagement on social media platforms such as Twitter, Instagram, and Facebook. The whole world is going to be connected by mobile devices with the evolution of networks in upcoming years. The advancement or the evolution of mobile network is evident from the increasing number of mobile devices, large connectivity, high data traffic, and increased widespread mobile applications. The massive MIMO technique is one of the important technologies for future 5G or 6G networks. The large number of antennas at BS is basically the expansion of conventional MIMO system which provide increased throughput and enhanced spectral efficiency. The essence of this technology lies in the integration of antennas, radios, and accessible spectrums to allow higher speed and capacity for the upcoming 5G era [2]. Considering its ability to boost throughput and spectral efficiency, massive MIMO is currently regarded as a technology that will be required for upcoming wireless standards [3], [4]. The crucial aspect lies in the substantial array gain achieved by massive MIMO through the utilization of a significant number of antennas [4]. One key technological advancement that makes 5G and advanced networks possible is massive MIMO. Massive MIMO and intelligent sensing systems are closely related because intelligent sensing systems mainly depend on 5G and beyond networks to operate.

Conventional multi-access methods for collecting data from multiple smart sensors are very unfeasible and result in decreased reliability, low data rates, and high latency. However, Massive MIMO is excellent at detecting data from several sensor transmissions at once, drastically cutting latency, and giving sensors higher data rates and more stable connections. This is due to its extensive beam forming and multiplexing capabilities. The real-time transmission of data gathered by smart sensors to central monitoring hubs is expected to be made possible by massive MIMO systems. This will enable a variety of applications, including intelligent highways, innovative buildings, autonomous vehicles, remote healthcare, smart grids, advanced antennas, and environmental monitoring.

Fig. 1 shows the number of connected devices and data traffic for each year from 2020 to 2030, and it is evident that both the number of devices per person and data traffic are steadily rising each year.

The authors acknowledge the support of the Hamdard University, Karachi, Pakistan, in facilitating this research.

First author is associated with Hamdard University, Karachi, Pakistan (e-mail: muhammad.ahsan@hamdard.edu.pk).

Second author is affiliated with Hamdard University, Karachi, Pakistan (e-mail: tayyab.ahmed@hamdard.edu.pk).

Third author is associated with Iqra University, Karachi, Pakistan (e-mail: sadiq.rehman@iqra.edu.pk).

Fourth author is associated with Hamdard University, Karachi, Pakistan (e-mail: halar.mustafa@hamdard.edu.pk).

DOI: 10.36244/ICJ.2025.2.12

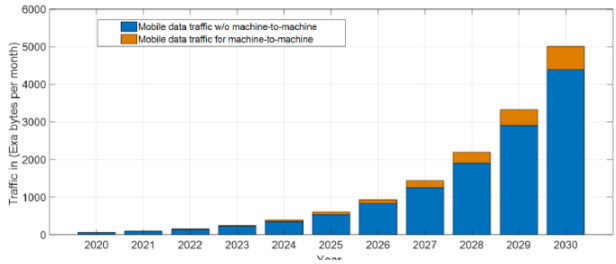


Fig. 1: Global mobile data traffic forecast for the period 2020 to 2030 [1].

Multiuser interference (MUI) inside the system is the main obstacle faced in the development of huge MIMO receivers. This challenge arises due to the simultaneous reception and transmission of users simultaneously, which results in severe degradation of overall system performance. A channel estimate method must reduce the MUI in order to fully utilize the massive MIMO system. A downlink massive MIMO system is taken into consideration in this study. To improve channel estimation performance, a constrained LS solution has been proposed. The outcomes demonstrate that the suggested constrained LS solutions outperform the unconstrained LS algorithm. The constrained algorithm is adopted from [5] for constrained channel estimation. Constrained least square channel estimation is a technique used to improve wireless communication system channel estimation accuracy while maintaining constraints pertaining to operational or physical characteristics of the system.

The rest of the paper is organized below. The most recent advancements in massive MIMO communication are covered in Section 2, recent trends in channel estimation of massive MIMO is discussed in Section 3, and the system model is explained in Section 4. Channel estimation is covered in Section 5, and the outcomes are covered in Section 6. Section 7 concludes the entire paper.

II. MASSIVE MIMO SYSTEMS

MIMO techniques are a crucial component of contemporary wireless networks, and in order to achieve significant improvements in both spectral and energy efficiency, they have been used more and more in recent years [26]. SISO techniques were common before MIMO was adopted; these systems could not reliably sustain a great number of users due to their low throughput. Many MIMO technologies, including single-user MIMO (SU-MIMO), have been introduced to meet the growing demand from multiple users [6], [7], MU-MIMO [8] and network MIMO [9], [10] were created. But even with these state-of-the-art technologies, the demands are too great to meet on their own. The exponential increase in wireless users in recent years has resulted in trillions of data that need to be controlled with efficiency and dependability. Additionally, the use of billions of Internet of Things (IoT) devices for smart energy, smart healthcare, and smart homes contributes to the rise in data traffic.

The 4G/LTE networks' existing MIMO technologies are inadequate to deal with this significant surge in traffic of data with the necessary speed and dependability. Massive MIMO technology is thus being investigated by the 5G network as a potential remedy for the problems brought on by the massive data traffic and the expanding user [3], [11]. The massive MIMO system has been studied extensively for its advantages [12]. One of the most important technologies for 5G and 6G is the massive MIMO communication technology. It is an advanced version of conventional MIMO, however in massive MIMO, thousands of antennas at the base station service tens of users [13]. In mm-Wave communication, antennas can be combined in a small area as compared to the microwave due to the small wavelength. Fig. 2 depicts the uplink and downlink massive MIMO system. Numerous antennas at the base station can create a directed beam for the intended user, allowing for great throughput at the user end with little interference to nearby users. When using massive MIMO instead of traditional MIMO systems, a high spectral efficiency is possible. Improved capacity and better signal quality are accomplished via numerous antennas, leading to sophisticated wireless communication systems [14]. The effect of rain on throughput is investigated in [27] and an AI-based solution is proposed to mitigate the effect of rain on throughput. The solar panel-based 3D array antennas is investigated in [28] for MIMO applications and no negative effects on the antenna system performance.

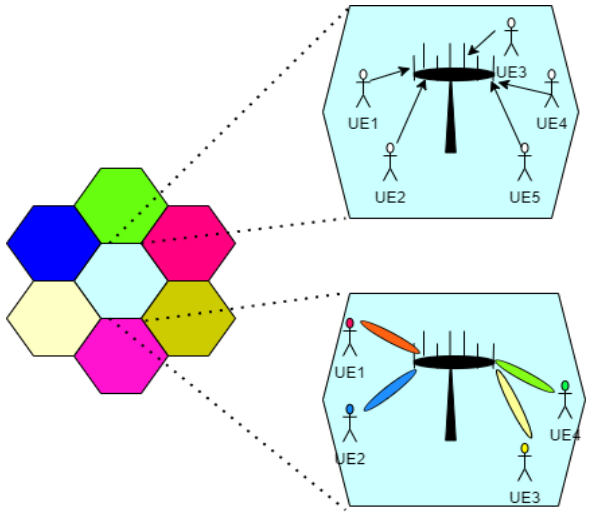


Fig.2: Massive MIMO uplink and downlink.

III. RECENT TRENDS IN MASSIVE MIMO CHANNEL ESTIMATION

Channel estimation is a technique to estimate the condition of wireless channels such as distortion, multipath effect and signal attenuation etc., the robust wireless communication system is only possible with effective channel estimation. The massive MIMO channel estimation recent trends is shown in Table 1.

TABLE I
CHANNEL ESTIMATION TECHNIQUES

Ref.	System Model	Channel Estimation Technique
[15]	Uplink system model with Gaussian mixture model (GMM).	Machine learning-based Channel Estimation.
[16]	Uplink training in flat Rayleigh fading channels.	Maximum likelihood-based MMSE channel estimator.
[17]	Uplink XL-MIMO OFDM communication.	Near-field XL-MIMO channel estimation schemes.
[18]	Massive MIMO system operating at millimeter waves using a lens array	Channel Estimation based on Deep Learning.
[19]	Downlink massive MIMO system	Compressive sensing-based channel estimation.
[20]	Down link massive MIMO	Channel estimation based on block iterative support detection.
[21]	Rayleigh fading Downlink channel massive MIMO system model.	There are two methods proposed: one makes use of a neural network with fully connected layers, and the other makes use of a CNN.
[22]	Narrowband flat block fading multiuser massive MIMO system.	Suggested a method to streamline CSI acquisition and decrease pilot overhead
[23]	One-bit mmWave massive MIMO	The Fisher information matrix (FIM) for these channel models.
[24]	millimeter-wave (mmWave) multiple-input multiple-output (MIMO) systems	Channel estimation approach using an iterative reweighting log-sum constraint.
[25]	Multiuser massive MIMO uplink system	Estimating Channels in Hybrid Massive MIMO Systems with Adaptive-Resolution ADCs.

IV. SYSTEM MODEL

In this paper, a downlink single-cell massive MIMO system with multiple base station antennas serving K single-antenna users simultaneously is examined given in Fig 3.

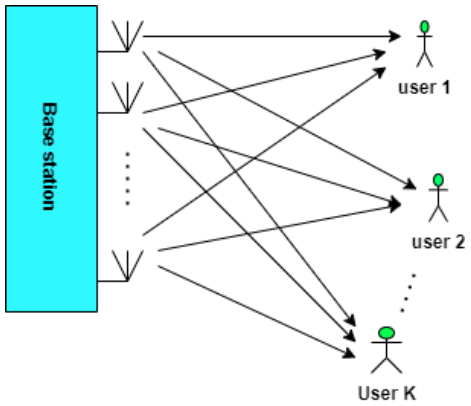


Fig. 3: System Model.

In this system model, antennas are arranged in an array linearly, which allows simplified design and implementation. The above arrangement is used to obtain enhanced beamforming. It will provide more coverage to the user available in the system in any direction.

The signal that was received at time i is shown below.

$$r(i) = \sum_{k=1}^{K} \boldsymbol{u}_k(i)\boldsymbol{h} + v(i) \tag{1}$$

where $\boldsymbol{u}_k(i) = [u_k(1), u_k(2), \cdots, u_k(M)]$, $\boldsymbol{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_M \end{bmatrix}$, and v is noise.

In (1), desired user, interference, and noise can be written as follows

$$r(i) = \underbrace{\boldsymbol{u}_1(i)\boldsymbol{h}}_{\text{desired}} + \underbrace{\sum_{\boldsymbol{k=2}}^{\boldsymbol{K}} \boldsymbol{u_k}(i)\boldsymbol{h}}_{\text{interference}} + \underbrace{\boldsymbol{v}(i)}_{\text{noise}} \tag{2}$$

N is the number of samples used for the channel estimation procedure, and the variables in (2) can be specified as follows.

$$\boldsymbol{y} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(N) \end{bmatrix}, \quad \boldsymbol{U}_k = \begin{bmatrix} \boldsymbol{u}_k(1) \\ \boldsymbol{u}_k(2) \\ \vdots \\ \boldsymbol{u}_k(N) \end{bmatrix}, \boldsymbol{v} = \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(N) \end{bmatrix}$$

Now, (2) can be written as follows

$$\boldsymbol{y} = \boldsymbol{U}_1\boldsymbol{h} + \sum_{k=2}^{K} \boldsymbol{U}_k\boldsymbol{h} + \boldsymbol{v} \tag{3}$$

Now, (3) can also be written as

$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{h} + \boldsymbol{v} \tag{4}$$

Where $\boldsymbol{U}$ is the sum of all $\boldsymbol{U}_k$

$$\boldsymbol{U} = \sum_{k=1}^{K} \boldsymbol{U}_k \tag{5}$$

V. CHANNEL ESTIMATION

Since the estimated channel primarily determines overall system performance and error-free sent data detection, it is an essential component of wireless communication systems. It is a well-known fact that the performance of any optimization goal improves by designing proper constraints on the optimization objective. This is due to the fact that the constraints restrict the search space for the optimization task and hence improves the convergence of the constrained optimization algorithm. Therefore, the channel estimate in multiuser massive MIMO system, constrained and un-constrained least square (LS) solution is designed. In this section LS based channel estimation technique is derived by considering different constraint such as when desired user pilots are known and when all the users pilot are known and results are compared to investigate the performance of each designed algorithms.

A. Unconstrained LS (Least Square) when all users' pilots are known

We can assess it as follows by using the MSE as the cost function to be minimized

$$J = E\left[\left\|\widehat{\boldsymbol{h}} - \boldsymbol{h}\right\|^2\right] \tag{6}$$

Estimated channel can be written as

$$\widehat{\boldsymbol{h}} = \boldsymbol{W}_{LS}\boldsymbol{y} \tag{7}$$

We can solve the norm and J expand as

$$J = E\left[\left(\widehat{\boldsymbol{h}} - \boldsymbol{h}\right)^H\left(\widehat{\boldsymbol{h}} - \boldsymbol{h}\right)\right] \tag{8}$$

Now substituting value of $\widehat{\boldsymbol{h}}$ from (7)

$$J = E[(\boldsymbol{W}_{LS}\boldsymbol{y} - \boldsymbol{h})^H(\boldsymbol{W}_{LS}\boldsymbol{y} - \boldsymbol{h})] \tag{9}$$

Solving transpose, we get

$$J = E\left[\left(\boldsymbol{y}^H{\boldsymbol{W}_{LS}}^H - \boldsymbol{h}^H\right)(\boldsymbol{W}_{LS}\boldsymbol{y} - \boldsymbol{h})\right] \tag{10}$$

Now substituting the value of $\boldsymbol{y}$ from equation (4)

$$J = E\left[\left\{(\boldsymbol{h}^H\boldsymbol{U}^H + \boldsymbol{v}^H){\boldsymbol{W}_{LS}}^H - \boldsymbol{h}^H\right\}\{\boldsymbol{W}_{LS}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{v}) - \boldsymbol{h}\}\right] \tag{11}$$

Solving ${\boldsymbol{W}_{LS}}^H$ and $\boldsymbol{W}_{LS}$ inside bracket we obtain

$$J = E\left[\left\{\boldsymbol{h}^H\boldsymbol{U}^H{\boldsymbol{W}_{LS}}^H + \boldsymbol{v}^H{\boldsymbol{W}_{LS}}^H - \boldsymbol{h}^H\right\}\{\boldsymbol{W}_{LS}\boldsymbol{U}\boldsymbol{h} + \boldsymbol{W}_{LS}\boldsymbol{v} - \boldsymbol{h}\}\right]$$

Applying property of trace, we get,

$$J = Tr\left(E\left[\{(\boldsymbol{W}_{LS}\boldsymbol{U} - \boldsymbol{I}_M)\boldsymbol{h} + \boldsymbol{W}_{LS}\boldsymbol{v}\}\left\{\boldsymbol{h}^H\left(\boldsymbol{U}^H{\boldsymbol{W}_{LS}}^H - \boldsymbol{I}_M\right) + \boldsymbol{v}^H{\boldsymbol{W}_{LS}}^H\right\}\right]\right) \tag{12}$$

Rearranging the above equation, we receive

$$J = Tr\left((\boldsymbol{W}_{LS}\boldsymbol{U} - \boldsymbol{I}_M)E[\boldsymbol{h}\boldsymbol{h}^H]\left(\boldsymbol{U}^H{\boldsymbol{W}_{LS}}^H - \boldsymbol{I}_M\right)\right) + Tr\left(\boldsymbol{W}_{LS}E[\boldsymbol{v}\boldsymbol{v}^H]{\boldsymbol{W}_{LS}}^H\right) \tag{13}$$

Since $E[\boldsymbol{h}\boldsymbol{h}^H] = \boldsymbol{I}_M$ and $E[\boldsymbol{v}\boldsymbol{v}^H] = \sigma_n^2\boldsymbol{I}_N$

Therefore,

$$J = Tr\left((\boldsymbol{W}_{LS}\boldsymbol{U} - \boldsymbol{I}_M)\boldsymbol{I}_M\left(\boldsymbol{U}^H{\boldsymbol{W}_{LS}}^H - \boldsymbol{I}_M\right)\right) + Tr\left(\boldsymbol{W}_{LS}\sigma_n^2\boldsymbol{I}_N{\boldsymbol{W}_{LS}}^H\right) \tag{14}$$

After simplification we obtain

$$J = Tr\left((\boldsymbol{W}_{LS}\boldsymbol{U} - \boldsymbol{I}_M)\left(\boldsymbol{U}^H{\boldsymbol{W}_{LS}}^H - \boldsymbol{I}_M\right)\right) + \sigma_n^2\boldsymbol{I}_N Tr\left(\boldsymbol{W}_{LS}{\boldsymbol{W}_{LS}}^H\right) \tag{15}$$

Rearranging the above equation, we get

$$J = Tr\left(\boldsymbol{W}_{LS}\boldsymbol{U}\boldsymbol{U}^H{\boldsymbol{W}_{LS}}^H\right) - Tr(\boldsymbol{W}_{LS}\boldsymbol{U}) - Tr\left(\boldsymbol{U}^H{\boldsymbol{W}_{LS}}^H\right) + Tr(\boldsymbol{I}_M) + \sigma_n^2\boldsymbol{I}_N Tr\left(\boldsymbol{W}_{LS}{\boldsymbol{W}_{LS}}^H\right) \tag{16}$$

We determine the aforementioned cost function's gradient with respect to the weight matrix and set it to zero in order to identify the best solution.

$$\frac{\partial J}{\partial {\boldsymbol{W}_{LS}}^H} = \boldsymbol{W}_{LS}\boldsymbol{U}\boldsymbol{U}^H - \boldsymbol{0} - \boldsymbol{U}^H + \boldsymbol{0} + \sigma_n^2\boldsymbol{I}_N\boldsymbol{W}_{LS} = \boldsymbol{0}$$

Finally, we get the solution is, therefore,

$$\boldsymbol{W}_{LS} = \boldsymbol{U}^H(\boldsymbol{U}\boldsymbol{U}^H + \sigma_n^2\boldsymbol{I}_N)^{-\boldsymbol{1}} \tag{17}$$

B. Constrained LS when only desired user's pilot is known

In this case, we assume that $\boldsymbol{u}_1$is known and all other $\boldsymbol{u}_k$'s are unknown.

We apply the following constrained optimization

$$\min J = E\left[\widetilde{\boldsymbol{h}}\widetilde{\boldsymbol{h}}^H\right] \text{ subject to } \boldsymbol{W}_{CLS1}\boldsymbol{U}_1 = \boldsymbol{I}_M$$

Where $\widetilde{\boldsymbol{h}} = \widehat{\boldsymbol{h}} - \boldsymbol{h}$

Before proceeding further, we define

$$z \triangleq \sum_{k=2}^{K} U_k h + v \tag{18}$$

Therefore, equation (3) becomes

$$\boldsymbol{y} = \boldsymbol{U}_1\boldsymbol{h} + \boldsymbol{z} \tag{19}$$

Consequently, the problem of limited optimization can be formulated as

$$\min J = E\left[(\widehat{\boldsymbol{h}} - \boldsymbol{h})(\widehat{\boldsymbol{h}} - \boldsymbol{h})^H\right] \text{ subject to } \boldsymbol{W}_{CLS1}\boldsymbol{u}_1 = I_M$$

$$\min J = E[(\boldsymbol{W}_{CLS1}\boldsymbol{y} - \boldsymbol{h})(\boldsymbol{W}_{CLS1}\boldsymbol{y} - \boldsymbol{h})^H] \tag{20}$$

Using the values of **y** and **z**, we get

$$\begin{aligned}\min J &= E[(\boldsymbol{W}_{CLS1}\boldsymbol{u}_1\boldsymbol{h} + \boldsymbol{W}_{CLS1}\boldsymbol{z} \\ &\quad - \boldsymbol{h})(\boldsymbol{W}_{CLS1}\boldsymbol{u}_1\boldsymbol{h} + \boldsymbol{W}_{CLS1}\boldsymbol{z} - \boldsymbol{h})^H] \\ &= E[(\boldsymbol{W}_{CLS1}\boldsymbol{z})(\boldsymbol{W}_{CLS1}\boldsymbol{z})^H] \text{since } \boldsymbol{W}_{CLS1}\boldsymbol{u}_1 = I_M \\ &= E\left[\boldsymbol{W}_{CLS1}\boldsymbol{z}\boldsymbol{z}^H{\boldsymbol{W}_{CLS1}}^H\right]\end{aligned} \tag{21}$$

Rearranging above equation, we receive

$$\min J = \boldsymbol{W}_{CLS1}\, E[\boldsymbol{z}\boldsymbol{z}^H]{\boldsymbol{W}_{CLS1}}^H$$

$$\min J = \boldsymbol{W}_{CLS1}\boldsymbol{R}_z{\boldsymbol{W}_{CLS1}}^H \tag{22}$$

The solution is well-known [5] and is given by

$$\boldsymbol{W}_{CLS1} = (\boldsymbol{U}_1^H\boldsymbol{R}_z^{-1}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1^H\boldsymbol{R}_z^{-1} \tag{23}$$

To find $\boldsymbol{R}_z$, we can evaluate it as

$$\boldsymbol{R}_z = E[\boldsymbol{z}\boldsymbol{z}^H]$$

$$\boldsymbol{R}_z = \boldsymbol{E}\left[\left(\sum_{k=2}^{K}\boldsymbol{U}_k\boldsymbol{h} + \boldsymbol{v}\right)\left(\sum_{k=2}^{K}\boldsymbol{U}_k\boldsymbol{h} + \boldsymbol{v}\right)^H\right] \tag{24}$$

$$\boldsymbol{R}_z = E\left[\sum_{k=2}^{K}\boldsymbol{U}_k\, E[\boldsymbol{h}\boldsymbol{h}^H]\sum_{k=2}^{K}\boldsymbol{U}^{\boldsymbol{H}}{}_k\right] + E[\boldsymbol{v}\boldsymbol{v}^H] \tag{25}$$

Applying estimates, we get

$$\boldsymbol{R}_z = \boldsymbol{E}\left[\left(\sum_{k=2}^{K}\boldsymbol{U}_k\right)I_M\left(\sum_{k=2}^{K}\boldsymbol{U}^{\boldsymbol{H}}{}_k\right)\right] + \sigma_n^2 I_N$$

$$\boldsymbol{R}_z = \sum_{k=2}^{K}E[\boldsymbol{U}_k\boldsymbol{U}^{\boldsymbol{H}}{}_k] + \sigma_n^2 I_N$$

$$\boldsymbol{R}_z = \sum_{k=2}^{K}\boldsymbol{R}_{uk} + \sigma_n^2 I_N \tag{26}$$

Where

$$\begin{aligned}\boldsymbol{R}_{uk} &= E[\boldsymbol{U}_k\boldsymbol{U}^{\boldsymbol{H}}{}_k] \\ &= E\left[\begin{bmatrix}\boldsymbol{u}_k(1)\\ \boldsymbol{u}_k(2)\\ \vdots\\ \boldsymbol{u}_k(N)\end{bmatrix}[\boldsymbol{u}^{\boldsymbol{H}}{}_k(1)\boldsymbol{u}^{\boldsymbol{H}}{}_k(2)\ \ldots\ \boldsymbol{u}^{\boldsymbol{H}}{}_k(N)]\right] \\ &= Tr(\boldsymbol{R}_k)I_N\end{aligned} \tag{27}$$

C. Constrained LS when all users' pilots are known

In this section, we designed another constrained LS solution for the channel estimation by utilizing the knowledge of all the users' pilots. For this purpose, we use the combined pilot matrix $\boldsymbol{U}$ to formulate the constrained LS problem as follows

$$\min J = E\left[\left\|\widehat{\boldsymbol{h}} - \boldsymbol{h}\right\|^2\right] \text{ subject to } \boldsymbol{W}_{CLS2}\boldsymbol{U} = I_M$$

$$J = E[(\boldsymbol{W}_{CLS2}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{v}) - \boldsymbol{h})(\boldsymbol{W}_{CLS2}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{v}) - \boldsymbol{h})^H] \tag{28}$$

Simplifying we get

$$J = E[(\boldsymbol{W}_{CLS2}\boldsymbol{v})(\boldsymbol{W}_{CLS2}\boldsymbol{v})^H] \tag{29}$$

Applying estimate, we get

$$J = \boldsymbol{W}_{CLS2}\boldsymbol{R}_v{\boldsymbol{W}_{CLS2}}^H, \text{ where } \boldsymbol{R}_v = \sigma_n^2 I_N \tag{30}$$

Finally, solution of equation is

$$\boldsymbol{W}_{CLS2} = \left(\boldsymbol{U}{\boldsymbol{R}_v}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^H{\boldsymbol{R}_v}^{-1} \tag{31}$$

VI. RESULTS AND ANALYSIS

Simulation of constrained and unconstrained channel estimation is carried out to verify our three solutions. The graphs, in Fig. 3, show the results for LS for following three situations:

(i) Constrained LS when only desired user pilot is known
(ii) Unconstrained LS when all users' pilots are known
(iii) Constrained LS when all users' pilots are known

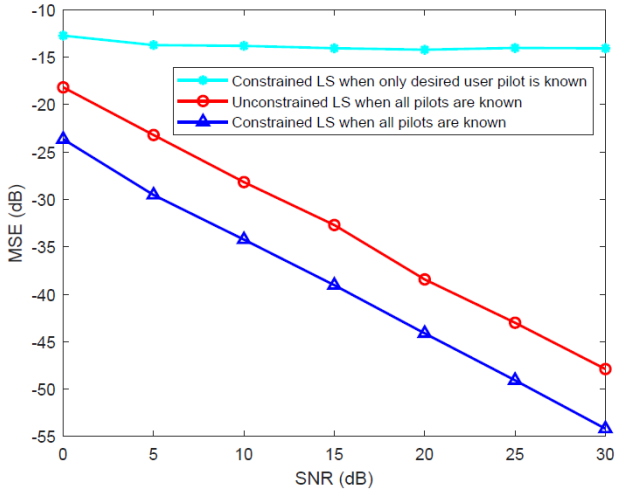


Fig. 4. Constrained LS and Unconstrained LS Algorithm.

In Fig 4, constrained LS with desired user pilot and constrained LS with all pilots are known compared with unconstrained LS when all pilots are known. The curves show that the constrained LS with all known pilots performs better than the unconstrained LS with all known pilots. Furthermore, the constrained LS with only a desired user-known pilot is poor as compared to other techniques. For example, at 30 dB SNR MSE of constrained LS, all pilots are known as -55 dB, and Constrained LS with the desired pilot is known as -13 dB. Additionally, the proposed channel estimation method is tested under various conditions, such as varying system user counts and antenna counts at the base station, and the outcomes are analyzed appropriately.

A. Effect of Number of Users

Different user conditions may have an impact on accuracy estimation because different user counts in the system may result in different amounts of interference. These results help to validate the adaptability and scalability of the proposed channel estimation algorithm.

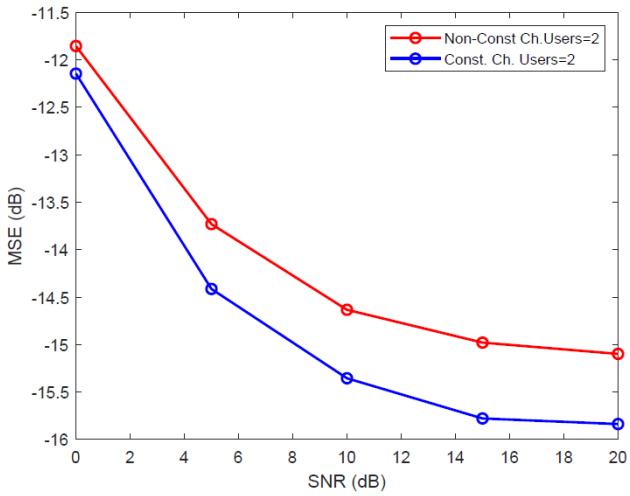


Fig.5: Constrained LS and Unconstrained LS Algorithm with 2 users.

In Fig. 5, Constrained LS and unconstrained LS are compared, and the system contains 2 users. In comparison to the unconstrained LS algorithm, the constrained LS channel estimation algorithm clearly performs better. For example, at SNR 20 dB MSE of the constrained LS algorithm is -12.7 dB, and the unconstrained LS algorithm is -11.7 dB.

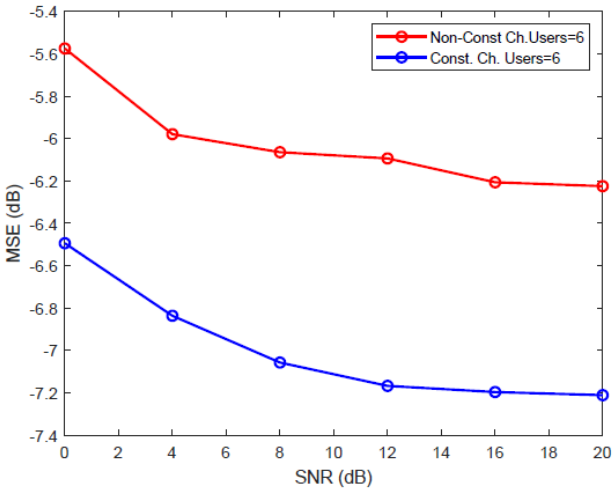


Fig. 6: Constrained LS and Unconstrained LS Algorithm with 6 users.

There are 6 users in the system, and Fig. 6 compares constrained and unconstrained LS. The findings demonstrate that system interference escalated with the number of users, resulting in a decrease in the performance of both methods. Results also show that the constrained LS algorithm performance is better than the unconstrained LS algorithm.

B. Effect of number of Antennas

This section varies the system's base station (BS) antenna size. In order to assess channel estimate performance, it is essential to simulate different numbers of transmission antennas because the number of antennas has a direct impact on spatial diversity, signal quality, and interference patterns. The proposed algorithm's scalability and efficiency to various massive MIMO scenarios may be assessed by evaluating multiple antenna designs. Results of constrained LS and unconstrained LS are compared and discussed accordingly.

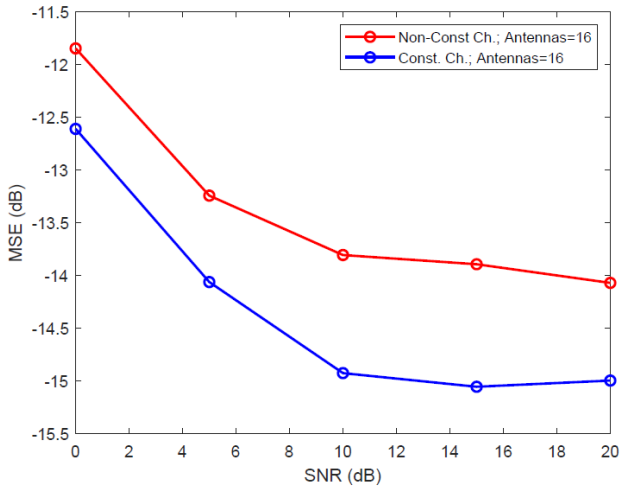


Fig. 7: Constrained LS and Unconstrained LS Algorithm BS antennas 16.

In Fig. 7, constrained LS and unconstrained LS algorithms are compared with BS having 16 antennas. Across all SNR values, the constrained LS algorithm continued to perform better than the unconstrained LS method. For example, at SNR 20 dB MSE of the constrained LS algorithm is -13.7 dB, and the unconstrained LS algorithm is -12.5 dB.

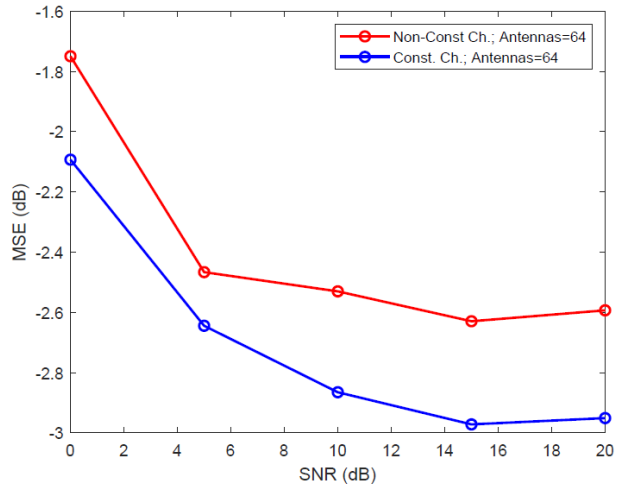


Fig. 8: Constrained LS and Unconstrained LS Algorithm BS antennas 64.

In Fig. 8, the performance of constrained LS and unconstrained LS algorithms are compared with 64 BS antennas. The results clearly show that the constrained LS algorithm gave much better performance. The findings also indicate that a base station's increased antenna count lowers channel estimation performance because it lengthens the channel path.

VII. CONCLUSION

A key component of 5G and beyond 5G communication networks is the large MIMO communication technology. The system's ability to efficiently detect user-transmitted data is primarily dependent on the base station's assessment of the channel. Channel estimation is always a challenging task and it became more difficult for large antenna systems. In order to improve the accuracy of the channel estimation, the constraint channel estimation technique is presented in this study. In the proposed channel estimation technique, two scenarios are considered, LS algorithm when only the desired user pilot is known, unconstrained LS when all users' pilots are known, and constrained LS when all users' pilots are known. Results show that the performance of constrained LS when all user pilots are known is better than the other two cases. Additionally, different numbers of users in the system and varied numbers of antennas at the base station are used to compare constrained versus unconstrained algorithms. The results show that the interference level of the system grows with the number of users, which degrades the channel estimation performance of both techniques. The advanced constraint modelling can be implemented as a future work along with the integration with machine learning to optimize the selection of constraints in real-time scenarios.

REFERENCES

[1] Karar AS, Falou ARE, Barakat JMH, Gürkan ZN, Zhong K. Recent Advances in Coherent Optical Communications for Short-Reach: Phase Retrieval Methods. *Photonics*. 2023; 10(3):308. DOI: 10.3390/photonics10030308

[2] F. Rusek, D. Persson, Buon Kiong Lau, E. G. Larsson, T. L. Marzetta, and F. Tufvesson, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013, DOI: 10.1109/MSP.2011.2178495.

[3] T. L. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Trans. Wirel. Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010, DOI: 10.1109/TWC.2010.092810.091092.

[4] X. Wu, N. C. Beaulieu, and D. Liu, "On Favorable Propagation in Massive MIMO Systems and Different Antenna Configurations," *IEEE Access*, pp. 1–1, 2017, DOI: 10.1109/ACCESS.2017.2695007.

[5] A. H. Sayed, Adaptive filters. Hoboken, N.J: Wiley-Interscience : IEEE Press, 2008.

[6] A. J. Paulraj and T. Kailath, "TRANSMISSION/DIRECTIONAL RECEPTION (DTDR)," 1994.

[7] G. J. Foschini and M. J. Gans, "On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas".

[8] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 60–67, Oct. 2004, DOI: 10.1109/MCOM.2004.1341262.

[9] S. Taruna and I. Kaur, "Performance analysis of MIMO for various antenna configurations," in *2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)*, CHENNAI, India: IEEE, Dec. 2013, pp. 90–93. DOI: 10.1109/ICGCE.2013.6823406.

[10] G. Caire, S. A. Ramprashad, and H. C. Papadopoulos, "Rethinking network MIMO: Cost of CSIT, performance analysis, and architecture comparisons," in *2010 Information Theory and Applications Workshop (ITA)*, La Jolla, CA, USA: IEEE, Jan. 2010, pp. 1–10. DOI: 10.1109/ITA.2010.5454094.

[11] N. H. M. Adnan, I. Md. Rafiqul, and A. H. M. Z. Alam, "Massive MIMO for Fifth Generation (5G): Opportunities and Challenges," in *2016 International Conference on Computer and Communication Engineering (ICCCE)*, Kuala Lumpur, Malaysia: IEEE, Jul. 2016, pp. 47–52. DOI: 10.1109/ICCCE.2016.23.

[12] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency," *Found. Trends® Signal Process.*, vol. 11, no. 3–4, pp. 154–655, 2017, DOI: 10.1561/2000000093.

[13] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," IEEE Commun. Mag., vol. 54, no. 2, pp. 114–123, Feb. 2016, DOI: 10.1109/MCOM.2016.7402270.

[14] Q. Hu, M. Zhang, and R. Gao, "Key Technologies in Massive MIMO," *ITM Web Conf.*, vol. 17, p. 01017, 2018, DOI: 10.1051/itmconf/20181701017.

[15] P. Su and Y. Wang, "Channel Estimation in Massive MIMO Systems Using a Modified Bayes-GMM Method," *Wirel. Pers. Commun.*, vol. 107, no. 4, pp. 1521–1536, Aug. 2019, DOI: 10.1007/s11277-019-06339-5.

[16] F. A. P. de Figueiredo, F. A. C. M. Cardoso, I. Moerman, and G. Fraidenraich, "Channel estimation for massive MIMO TDD systems assuming pilot contamination and flat fading," *EURASIP J. Wirel. Commun. Netw.*, vol. 2018, no. 1, p. 14, Dec. 2018, DOI: 10.1186/s13638-018-1021-9.

[17] M. Cui and L. Dai, "Channel Estimation for Extremely Large-Scale MIMO: Far-Field or Near-Field?" arXiv, Jan. 18, 2022. Accessed: Oct. 14, 2023. [Online]. Available: http://arxiv.org/abs/2108.07581

[18] R. Zhang, W. Tan, W. Nie, X. Wu, and T. Liu, "Deep Learning-Based Channel Estimation for mmWave Massive MIMO Systems in Mixed-ADC Architecture," *Sensors*, vol. 22, no. 10, p. 3938, May 2022, DOI: 10.3390/s22103938.

[19] Z. Albataineh, K. Hayajneh, H. Bany Salameh, C. Dang, and A. Dagmseh, "Robust massive MIMO channel estimation for 5G networks using compressive sensing technique," *AEU – Int. J. Electron. Commun.*, vol. 120, p. 153197, Jun. 2020, DOI: 10.1016/j.aeue.2020.153197.

[20] W. Shen, L. Dai, Y. Shi, Z. Gao, and Z. Wang, "Massive MIMO channel estimation based on block iterative support detection," in *2016 IEEE Wireless Communications and Networking Conference*, Doha, Qatar: IEEE, Apr. 2016, pp. 1–6. DOI: 10.1109/WCNC.2016.7564735.

[21] H. Hirose, T. Ohtsuki, and G. Gui, "Deep Learning-Based Channel Estimation for Massive MIMO Systems With Pilot Contamination," *IEEE Open J. Veh. Technol.*, vol. 2, pp. 67–77, 2021, DOI: 10.1109/OJVT.2020.3045470.

[22] I. Khan et al., "A Robust Channel Estimation Scheme for 5G Massive MIMO Systems," Wirel. Commun. Mob. Comput., vol. 2019, pp. 1–8, Aug. 2019, DOI: 10.1155/2019/3469413.

[23] S. Rao, A. Mezghani, and A. L. Swindlehurst, "Channel Estimation in One-Bit Massive MIMO Systems: Angular Versus Unstructured Models," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 5, pp. 1017–1031, Sep. 2019, DOI: 10.1109/JSTSP.2019.2933163.

[24] A. Zhang, W. Cao, P. Liu, J. Sun, and J. Li, "Channel Estimation for MmWave Massive MIMO With Hybrid Precoding Based on Log-Sum Sparse Constraints," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 68, no. 6, pp. 1882–1886, Jun. 2021, DOI: 10.1109/TCSII.2020.3041230.

[25] Y. Wang, X. Chen, Y. Cai, B. Champagne, and L. Hanzo, "Channel Estimation for Hybrid Massive MIMO Systems with Adaptive-Resolution ADCs." arXiv, Dec. 31, 2021. Accessed: Nov. 02, 2023. [Online]. Available: http://arxiv.org/abs/2112.15419

[26] Ammar Al-Adhami, Yasir Al-Adhami, and Taha A. Elwi, "A 3D Antenna Array based Solar Cell Integration for Modern MIMO Systems", *Infocommunications Journal*, Vol. XV, No 4, December 2023, pp. 10–16., DOI: 10.36244/ICJ.2023.4.2

[27] A. M. Al-Saegh et al., "AI-Based Investigation and Mitigation of Rain Effect on Channel Performance With Aid of a Novel 3D Slot Array Antenna Design for High Throughput Satellite System," in *IEEE Access*, vol. 12, pp. 29 926–29 939, 2024, DOI: 10.1109/ACCESS.2024.3368829.

[28] Al-Adhami, A., Al-Adhami, Y. and Elwi, T.A. (2023) 'A 3D antenna array based solar cell integration for modern MIMO Systems', *Infocommunications Journal*, 15(4), pp. 10–16. DOI: 10.36244/icj.2023.4.2.

Muhammad Ahsan Shaikh received the B.Eng. degree (Hons) in Electronics majoring in telecommunication from Multimedia University, Cyberjaya, Malaysia in 2014, and the M.E and Ph.D. in Electrical Engineering from Hamdard University, Karachi, Pakistan in 2017 and 2024 respectively. He has served as a research assistant at wireless communication lab, Multimedia University, Malaysia in 2014. He is also currently serving as an Assistant Professor at Faculty of Engineering Sciences and Technology, Hamdard University, Karachi. He is also associated with next generation mobile networks research cluster, Hamdard University. His research interests include beamforming design of MIMO NOMA and signal processing for wireless communications.

Tayyab Ahmed Shaikh is currently working as Assistant Professor at Faculty of Engineering Sciences and Technology, Hamdard University, Karachi, Pakistan. He has done BE and ME in Computer Engineering from NED University of Engineering and Technology, Karachi, Pakistan in the years 2005 and 2010 respectively. He earned the PhD degree in Electronic Engineering from Hamdard University, Karachi, Pakistan in the year 2021. His research interest includes Optimization, Signal Processing and Control, Artificial Intelligence, Computer Architecture.

Sadiq Ur Rehman is an associate professor in computer science, FEST, Iqra University, Karachi, Pakistan. He was associated with Hamdard University from April 2016 to April 2025 and also completed Ph.D. in Electrical Engineering (Specialization in Communication System and Networks) from the same University. He has done MS in Computer Science and Communication Engineering from University Duisburg-Essen, Germany in 2015 and BE in Electronics Engineering from Usman Institute of Technology (UIT), Karachi, Pakistan in 2008 respectively. His current research interests include Sensor Networks, Wireless Body Area Network and IoT.

Halar Mustafa is currently working as a lecturer and coordinator of the electrical engineering technology program at Hamdard University. He has received his Bachelor of Engineering in Telecom Engineering and Masters of Engineering in Electrical Engineering from Hamdard University, Pakistan, Currently, he is a Ph.D. scholar of Electrical Engineering at Hamdard University, Pakistan. He has received the prestigious, four years of fully funded, Prime Minister National ICT Scholarship for undergraduate studies, from the Ministry of IT, Government of Pakistan. Further, he has received a merit-based scholarship during his master's degree. He has also obtained several funding from different agencies for his ongoing projects. His research interests are wireless communication and smart grid technology.

Guidelines for our Authors

Format of the manuscripts

Original manuscripts and final versions of papers should be submitted in IEEE format according to the formatting instructions available on

https://journals.ieeeauthorcenter.ieee.org/
Then click: "IEEE Author Tools for Journals"
- "Article Templates"
- "Templates for Transactions".

Length of the manuscripts

The length of papers in the aforementioned format should be 6-8 journal pages.
Wherever appropriate, include 1-2 figures or tables per journal page.

Paper structure

Papers should follow the standard structure, consisting of *Introduction* (the part of paper numbered by "1"), and *Conclusion* (the last numbered part) and several *Sections* in between.
The Introduction should introduce the topic, tell why the subject of the paper is important, summarize the state of the art with references to existing works and underline the main innovative results of the paper. The Introduction should conclude with outlining the structure of the paper.

Accompanying parts

Papers should be accompanied by an *Abstract* and a few *Index Terms (Keywords)*. For the final version of accepted papers, please send the short cvs and *photos* of the authors as well.

Authors

In the title of the paper, authors are listed in the order given in the submitted manuscript. Their full affiliations and e-mail addresses will be given in a footnote on the first page as shown in the template. No degrees or other titles of the authors are given. Memberships of IEEE, HTE and other professional societies will be indicated so please supply this information. When submitting the manuscript, one of the authors should be indicated as corresponding author providing his/her postal address, fax number and telephone number for eventual correspondence and communication with the Editorial Board.

References

References should be listed at the end of the paper in the IEEE format, see below:

a) Last name of author or authors and first name or initials, or name of organization
b) Title of article in quotation marks
c) Title of periodical in full and set in italics
d) Volume, number, and, if available, part
e) First and last pages of article
f) Date of issue
g) Document Object Identifier (DOI)

[11] Boggs, S.A. and Fujimoto, N., "Techniques and instrumentation for measurement of transients in gas-insulated switchgear," IEEE Transactions on Electrical Installation, vol. ET-19, no. 2, pp.87–92, April 1984. DOI: 10.1109/TEI.1984.298778

Format of a book reference:

[26] Peck, R.B., Hanson, W.E., and Thornburn, T.H., Foundation Engineering, 2nd ed. New York: McGraw-Hill, 1972, pp.230–292.

All references should be referred by the corresponding numbers in the text.

Figures

Figures should be black-and-white, clear, and drawn by the authors. Do not use figures or pictures downloaded from the Internet. Figures and pictures should be submitted also as separate files. Captions are obligatory. Within the text, references should be made by figure numbers, e.g. "see Fig. 2."
When using figures from other printed materials, exact references and note on copyright should be included. Obtaining the copyright is the responsibility of authors.

Contact address

Authors are requested to submit their papers electronically via the following portal address:

https://www.ojs.hte.hu/infocommunications_journal/about/submissions

If you have any question about the journal or the submission process, please do not hesitate to contact us via e-mail:

Editor-in-Chief: Pál Varga – pvarga@tmit.bme.hu

Associate Editor-in-Chief:

József Bíró – biro@tmit.bme.hu

László Bacsárdi – bacsardi@hit.bme.hu

IEEE International Conference on Communications

24–28 May 2026 // Glasgow, Scotland, UK

CALL FOR PAPERS

The 2026 IEEE International Conference on Communication (ICC) 2026 will be held in the Scottish Event Campus and the Crowne Plaza Hotel in Glasgow, Scotland from 24 – 28 May 2026. Themed "*Connected World for Sustainable Future,*" this flagship conference of the IEEE Communications Society will feature a comprehensive high-quality technical program including 12 symposia and a variety of tutorials and workshops. IEEE ICC 2026 will also include an attractive industry program aimed at practitioners, with keynotes and panels from prominent research, industry and government leaders, business and industry panels, and technological exhibits. This conference will be fully in-person.

SYMPOSIA

- Cognitive Radio & AI-Enabled Networks
- Communication & Information System Security
- Communication QoS, Reliability & Modelling
- Communications Software & Multimedia
- Communication Theory
- Green Communication Systems & Networks
- IoT & Sensor Networks
- Mobile & Wireless Networks
- Next-Generation Networking & Internet
- Optical Networks & Systems
- Signal Processing for Communications
- Wireless Communications

SELECTED AREAS IN COMMUNICATIONS

- Aerial Communications
- Backhaul and Fronthaul Communications
- Big Data
- Cloud Computing and Networking
- eHealth
- Integrated Sensing and Communications
- Machine Learning for Communications
- Molecular, Biological, and Multi-Scale Communications
- Next Generation Multiple Access
- Quantum Communications and Computing
- Reconfigurable Intelligent Surfaces and Holographic MIMO
- Satellite and Space Communications
- Social Networks

ORGANIZING COMMITTEE

General Chair
Tariq. S. Durrani, Univ. Strathclyde, UK

Executive Chairs
Muhammad Ali Imran, Univ. Glasgow, UK
Muhammad Zeeshan Shakir, UWS, UK

TPC Chair
Marco Di Renzo,
King's College London, UK
CNRS & CentraleSupélec, France

TPC Co-Chairs
Leila Musaivan, Univ. Essex, UK
Tony Q. S. Quek, STUD, Singapore

Keynote Chairs
Urbashi Mitra, USC, USA
Lajos Hanzo, Univ. Southampton, UK

Operations Chair
Qammer Abbasi,
Univ. of Glasgow, UK

Finance Chair
Ray Dent, NGS, UK

IF&E Chairs
Majid Butt, Nokia, USA
Jacquiline Redmond, CENSIS, UK

GIMS Advisor
Nordin Ramli,
MIMOS, Malaysia

GITC Advisor
Yacine Ghamri-Doudane,
Univ., La Rochelle, France

IEEE ComSoc UK & I Chair
Nauman Aslam, Northumbria Univ., UK

Awards Co-chairs
Angel Lozano,
Univ. Pompeu Fabra, Spain
Lin Cai, Univ. Victoria, Canada

Tutorial Program Co-chairs
Nizar Zorba, Qatar Univ., Qatar
Kuljeet Kaur, ETS Montréal, Canada

Workshop Program Co-chairs
Ramon Aguero, Univ. Cantabria Spain
Miguel A. Dajer, Futurewei, USA
Cheng-Xiang Wang, Southeast Univ., China

Travel Grants Co-chairs
Matthew Valenti,
West Virginia Univ., USA
Virginia Pilloni, Univ. Cagliari, Italy
Anastassia Gharib,
Princess Sumaya Univ., Jordan

Publications Co-chairs
James Irvine, Univ. Strathclyde, UK
Zeynep Gurkas-Aydin,
Istanbul Univ.-Cerrahpasa, Turkey

Publicity Chairs
Angela Zhang,
The Chinese Univ. Hong Kong
Aryan Kaushik, Manchester MET, UK
Yessica Saez,
Technol. Univ. Panama, Panama
Marwa Chafii,
New York Univ., Abu Dhabi

IMPORTANT DATES

Paper Submission:
29 September 2025

Acceptance Notification:
12 January 2026

Camera-Ready:
13 February 2026

Panel Proposals:
1 December 2025

Workshop Proposals:
7 September 2025

Tutorial Proposals:
22 September 2025

IEEE ICC®

icc2026.ieee-icc.org

IEEE ComSoc | IEEE

SCIENTIFIC ASSOCIATION FOR INFOCOMMUNICATIONS

Who we are

Founded in 1949, the Scientific Association for Infocommunications (formerly known as Scientific Society for Telecommunications) is a voluntary and autonomous professional society of engineers and economists, researchers and businessmen, managers and educational, regulatory and other professionals working in the fields of telecommunications, broadcasting, electronics, information and media technologies in Hungary.

Besides its 1000 individual members, the Scientific Association for Infocommunications (in Hungarian: HÍRKÖZLÉSI ÉS INFORMATIKAI TUDOMÁNYOS EGYESÜLET, HTE) has more than 60 corporate members as well. Among them there are large companies and small-and-medium enterprises with industrial, trade, service-providing, research and development activities, as well as educational institutions and research centers.

HTE is a Sister Society of the Institute of Electrical and Electronics Engineers, Inc. (IEEE) and the IEEE Communications Society.

What we do

HTE has a broad range of activities that aim to promote the convergence of information and communication technologies and the deployment of synergic applications and services, to broaden the knowledge and skills of our members, to facilitate the exchange of ideas and experiences, as well as to integrate and harmonize the professional opinions and standpoints derived from various group interests and market dynamics.

To achieve these goals, we…

- contribute to the analysis of technical, economic, and social questions related to our field of competence, and forward the synthesized opinion of our experts to scientific, legislative, industrial and educational organizations and institutions;
- follow the national and international trends and results related to our field of competence, foster the professional and business relations between foreign and Hungarian companies and institutes;
- organize an extensive range of lectures, seminars, debates, conferences, exhibitions, company presentations, and club events in order to transfer and deploy scientific, technical and economic knowledge and skills;
- promote professional secondary and higher education and take active part in the development of professional education, teaching and training;
- establish and maintain relations with other domestic and foreign fellow associations, IEEE sister societies;
- award prizes for outstanding scientific, educational, managerial, commercial and/or societal activities and achievements in the fields of infocommunication.

Contact information

President: **FERENC VÁGUJHELYI** • *elnok@hte.hu*

Secretary-General: **GÁBOR KOLLÁTH** • *kollath.gabor@hte.hu*

Operations Director: **PÉTER NAGY** • *nagy.peter@hte.hu*

Address: H-1051 Budapest, Bajcsy-Zsilinszky str. 12, HUNGARY, Room: 502

Phone: +36 1 353 1027

E-mail: *info@hte.hu*, Web: *www.hte.hu*