In the Name of God

Journal of
Information Systems & Telecommunication
Vol. 14, No.3, July-September 2026, Serial Number 55

Research Institute for Information and Communication Technology
Iranian Association of Information and Communication Technology
Affiliated with: Academic Center for Education, Culture and Research (ACECR)

Manager-in-Charge: Dr. Ali Mokhtarani, ACECR, Iran
Editor-in-Chief: Dr. Masoud Shafiee, Amir Kabir University of Technology, Iran

Editorial Board
Dr. Abdolali Abdipour, Professor, Amirkabir University of Technology, Iran
Dr. Ali Akbar Jalali, Professor, Iran University of Science and Technology, Iran
Dr. Alireza Montazemi, Professor, McMaster University, Canada
Dr. Ali Mohammad-Djafari, Associate Professor, Le Centre National de la Recherche Scientifique (CNRS), France
Dr. Hamid Reza Sadegh Mohammadi, Associate Professor, ACECR, Iran
Dr. Mahmoud Moghavvemi, Professor, University of Malaya (UM), Malaysia
Dr. Mehrnoush Shamsfard, Associate Professor, Shahid Beheshti University, Iran
Dr. Omid Mahdi Ebadati, Associate Professor, Kharazmi University, Iran
Dr. Rahim Saeidi, Assistant Professor, Aalto University, Finland
Dr. Ramezan Ali Sadeghzadeh, Professor, Khajeh Nasireddin Toosi University of Technology, Iran
Dr. Sha'ban Elahi, Professor, Vali-e-asr University of Rafsanjan, Iran
Dr. Shohreh Kasaei, Professor, Sharif University of Technology, Iran
Dr. Habibollah Asghari, Associate Professor, ACECR, Iran
Dr. Zabih Ghasemlooy, Professor, Northumbria University, UK
Dr. Hadi Aliakbarian, Associate Professor, Khajeh Nasireddin Toosi University of Technology, Iran

Executive Editor: Dr. Fatemeh Kheirkhah
Executive Manager: Mahdokht Ghahari
Print ISSN: 2322-1437
Online ISSN: 2345-2773
Publication License: 91/13216
Editorial Office Address: No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,
P.O.Box: 13145-799 Tel: (+9821) 88930150 Fax: (+9821) 88930157
E-mail: info@jist.ir , infojist@gmail.com
URL: jist.acecr.org

Indexed by:
- SCOPUS www. Scopus.com
- Islamic World Science Citation Center (ISC) www.isc.gov.ir
- Directory of open Access Journals (DOAJ) www.Doaj.org
- Scientific Information Database (SID) www.sid.ir
- Regional Information Center for Science and Technology (RICeST) www.ricest.ac.ir
- Magiran www.magiran.com

Publisher:
Iranian Academic Center for Education, Culture and Research **(ACECR)**

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Telecommunication Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

(A-Z)

- Alaeiyan, Mohammad Hadi, K.N. Toosi University of Technology, Tehran, Iran
- Asghari, Habibollah, ACECR, Tehran, Iran
- Azarkasb, Seyed Omid, K.N. Toosi University of Technology, Tehran, Iran
- Agarwal, Parul, Jamia Hamdard University, New Delhi, India
- Abdulnabi, Saif, University of Kufa, Kufa, Iraq
- Bagheri, Ali, Emamhossein University, Tehran, Iran
- Dehghani, Maryam, Malek Ashtar University, Tehran, Iran
- Dehghani, Mohammad, Modares University, Tehran, Iran
- Ebadati, Omid Mahdi, Kharazmi University,Tehran, Iran
- Farsi, Hassan, University of Birjand, South Khorasan, Iran
- Farrokhbakht Foumani, Mehdi, Islamic Azad University Fouman and Shaft, Gilan, Iran
- Ghasemzadeh, Mohammad, Yazd University, Yazd, Iran
- Ghasemi, Jamal, University of Mazandaran, Babolsar, Iran
- Goodarzi, Marjan, Research Institute for Information and Communication Technology, Tehran, Iran
- Jafargholi, Amir, EPFL Swiss Federal Technology Institute of Lausanne, Lausanne, Switzerland
- Kashef, Seyed Sadra, Urmia University, Urmia, Iran
- Kheirkhah, Fatemeh, ACECR, Tehran, Iran
- Kazerouni, Morteza, Malek-Ashtar University of Technology, Tehran, Iran
- Kolahkaj, Maral, Islamic Azad University, Karaj Branch, Iran
- Mansouri, Azadeh, Kharazmi University,Tehran, Iran
- Mirroshandel, Seyed Abolghasem, University of Guilan, Rasht, Iran
- Molazadeh,Amir Hosein, K. N. Toosi University of Technology, Tehran, Iran
- Mohammadzadeh, Sajjad, University of Birjand, South Khorasan, Iran
- Mirzaei, Abbas, Islamic Azad University, Ardabil, Iran
- Nangir, Mahdi, University of Tabriz, Tabriz, Iran
- Najimi, Maryam, University of Science and Technology of Mazandaran, Behshahr, Iran
- Razavi, Seyyed Mohammad, University of Birjand, South Khorasan Province, Iran
- Rasi, Habib, Shiraz University of Technology, Shiraz, Iran
- Riahi, Noushin, Alzahra University, Tehran, Iran
- Sadeghzadeh, Ramezan Ali, K. N. Toosi University of Technology, Tehran, Iran
- Saraeian, Shideh, Islamic Azad University, Gorgan Branch, Iran
- Soryani, Mohsen, University of science and technology, Tehran, Iran
- Sobuti, Saman, Islamic Azad University of keman, kerman, Iran
- Soleimani, Ali, Islamic Azad University, Malard Branch, Iran
- Shabani, Mehraneh, Mazandaran University of Medical Sciences, Sari, Iran
- Tourani, Mahdi, University of Birjand, South Khorasan, Iran
- Yaghoobi, Kaebeh, Ale Taha Institute of Higher Education, Tehran, Iran
- Zerehsaz, Mohammad, Kharazmi University,Tehran, Iran

Table of Contents

- **A Hybrid Deep Learning Modle Based on BLSTM-GRU for Enhanced Call Volume Forcasting in Call Centers 182**

Armaghan Nikfar, Sasan Harifi and Javad Mohammadzadeh

- **Neural Risk-Adaptive Access Control: A Behavior-Context Hybrid Framework for Adaptive Authorization in Dynamic Environments 196**

Donya Piriaei, Afshin Rezakhani, Hamid Haj Seyyedjavadi and Leila Rikhtechi

- **Fusion Based Deep Learning Framework Detection Using an Optimized for Cardiovascular Disease Squeeze Net Mode ..... .214**

N Shanmuga Sundari

- **Soft and Safe Network Slicing with Hierarchical Multi-Agent Reinforcement Learning 225**

Zeinab Sasan and Siavash Khorsandi

- **Assessing the Usability of a Comprehensive Research Management System: A Mixed-Use Study in Iran 238**

Mehdi Beyrami Jam, Mohammad Sadegh Sohrabi, Mohammad Babamiri and Arezoo Farhadi

- **A Hybrid Entity-Centric Approach to Persian Pronoun Resolution with BERT Enhancement 245**

Hassan Haji Mohammadi, Alireza Talebpoour, Ahmad Mahmoudi Aznaveh and Saman Yazdani

- **Hybrid Embedding and GPT-5 Instruction-Augmented Framework for Arabic Hate Speech Detection 268**

Anas Khudhur Abbas Al Juboori and Heshaam Faili

Journal of Information Systems and Telecommunication
Vol.14, No.4, July-September 2026, 182-195
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

A Hybrid Deep Learning Model Based on BLSTM-GRU for Enhanced Call Volume Forecasting in Call Centers

Armaghan Nikfar[1], Sasan Harifi[1*], Javad Mohammadzadeh[1]

[1].Department of Computer Engineering, Ka.C., Islamic Azad University, Karaj, Iran

Received: 05 Feb 2026/ Revised: 04 Sep 2026/ Accepted: 23 Sep 2026

Abstract

This research aims to develop a forecasting model for call center call volume by incorporating characteristics unique to each country and call center environment. Using historical call data, calendar events, and other relevant input variables, the goal is to create a framework that is capable of accurately simulating and predicting the volume of incoming calls over specified time periods. For this purpose, we employ advanced recurrent neural network architectures, including MLP, RNN, LSTM, and a hybrid BLSTM-GRU model, to analyze temporal patterns in call data. The BLSTM-GRU model, in particular, uses two-way and gating processing mechanisms to more effectively capture complex temporal dependencies. The dataset consists of daily call volume over a 4-year period along with text features formatted in Excel for ease of training and evaluation. Furthermore, the forecasting framework integrates a comparison mechanism that compares the predicted call volume against the actual values to assess the accuracy and robustness of the model. Experimental results show that the BLSTM-GRU model outperforms standard architectures, achieving a forecasting accuracy of 94.12%, whereas the optimized single LSTM baseline reaches 81.35%. To increase operational efficiency, the framework also generates shift schedules that comply with call center standards and prevent call queues by recommending the required number of agents. The superior performance of the proposed BLSTM-GRU model highlights its potential application in various contact centers around the world and provides a scalable solution for intelligent contact center management.

Keywords: Long Short-Term Memory (LSTM); Call Center; BLSTM-GRU; Hybrid Deep Learning; Prediction.

1- Introduction

Accurately forecasting call volume in call centers is a critical operational challenge that directly impacts service quality and staffing efficiency. According to Hussain et al [1], predicting traffic flow from live data streams necessitates the development of robust data-driven models. This task is inherently challenging due to the complex, dynamic nature of traffic patterns and their temporal dependencies. A key difficulty lies in accurately capturing the diverse and often unpredictable variations in traffic behavior. While early efforts relied on statistical methods and shallow neural networks with limited success, recent advances in deep learning have demonstrated significant potential, offering more accurate and reliable predictions.

To enhance the understanding of spatial and temporal dependencies in traffic data, some researchers have explored traditional machine learning techniques such as Support Vector Machines (SVM). For instance, Feng et al [32] proposed an Adaptive Multi-Kernel SVM with Spatial–Temporal Correlation (AMSVM-STC) for short-term traffic flow prediction. This approach focuses on modeling the inherent nonlinearity and randomness of traffic behavior at specific Points of Interest (POIs) by leveraging probabilistic modeling techniques. Although these methods offer valuable insights, they often struggle to scale with complex temporal sequences compared to modern deep learning frameworks.

Convolutional neural networks (CNNs) are another popular approach for traffic prediction, known for their ability to model spatial features. Ma et al. [3] proposed a CNN-based method that learns traffic as images and predicts large-scale, network-wide traffic speed with high accuracy.

While the above approaches focus on general traffic forecasting using pre-built machine learning and deep learning models, our research addresses a specific and operationally critical area, which is call volume forecasting in call centers. Previously, a BLSTM model [4] has been trained and fitted with historical call data from an Iranian

✉ **Sasan Harifi**
harifi@iau.ac.ir

call center (Shatel) and incorporated specific features of the Persian calendar, such as national and religious holidays. The model predicted both hourly call volume and the optimal number of agents per shift. It incorporated micro-level operational factors, including the average number of calls per minute per agent, the probability of agent unavailability due to downtime, and the average call handling time per shift. By integrating these factors, this model can proactively suggest accurate staffing schedules to reduce queue formation and improve customer satisfaction.

We propose a novel hybrid architecture that combines bidirectional LSTM (BLSTM) and Gated Recurrent Unit (GRU) to further enhance prediction accuracy and robustness. The goal of this integration is to enhance the long-term dependency absorption capability of BLSTM and the lightweight and efficient structure of GRU, resulting in a more robust model that is able to more effectively address complex call center traffic patterns. We also want to use the Shatel call center shift scheduling policies to schedule experts on shifts so that the improved model can output a year of incoming calls with the required experts, both respecting the probabilistic principles and considering the expert shift scheduling policies. The main contributions of this paper are a novel hybrid BLSTM-GRU for call volume forecasting; integration of seven culturally specific calendar features (Nowruz, Ramadan, Muharram, etc.); a probabilistic agent scheduling mechanism that accounts for long-call variability; and a generalizable architecture adaptable to any country's calendar.

The remainder of this paper is organized as follows: Section 2 provides a review of related works in the field of call volume forecasting and the application of deep learning models in time series prediction. Section 3 introduces the proposed hybrid BLSTM-GRU approach. Section 4 presents the experimental results. Finally, Section 5 concludes the paper by summarizing the main findings and suggesting future research directions aimed at improving the scalability and adaptability of the proposed framework for global call center applications.

2- Related Works

Accurate forecasting of incoming call volumes in call centers plays a crucial role in ensuring service quality, reducing customer wait times, and optimizing workforce allocation. An overloaded call center, caused by underestimation of call traffic, leads to long queues, customer dissatisfaction, and service delays. Conversely, overestimation results in unnecessary staffing costs and resource waste. Thus, predictive modeling has become an essential strategy for modern call centers seeking to balance operational efficiency with high service-level standards. According to Aattouri et al [15], predicting incoming call flows enables better organization within call centers. It also reduces the number of lost calls and improves customer satisfaction. Also, Kanthanathan et al [16] said that a call center, customer service agents (CSAs) provide valuable products or information to the provider. A key requirement in a call center is to ensure customer satisfaction by having enough customer service agents (CSAs) to support incoming calls, while avoiding excess staffing to control costs.

Few studies have been conducted on the forecasting of call arrivals. Nikfar and Mohammadzadeh [4] proposed a deep learning-based framework using a Bidirectional Long Short-Term Memory (BLSTM) network to predict hourly call volumes at Shatel, a leading Iranian Internet Service Provider. The model was trained on four years of historical call data and uniquely incorporated temporal features specific to the Persian calendar, including official holidays (e.g., Nowruz), religious events, and weekends. By dynamically encoding these culturally significant features, the model was capable of adjusting to different calendar years without requiring manual feature engineering for each prediction cycle. The BLSTM model captured both forward and backward temporal dependencies in the data, enabling it to identify complex sequential patterns in call traffic.

Beyond predicting the volume of calls, the model extended its functionality to support human resource planning by estimating the number of agents required per hour. This was achieved through the integration of multiple operational factors such as average call handling time, probability of agent availability during breaks, and historical workload fluctuations. These elements were incorporated as input features, allowing the model to generate real-time staffing recommendations that aligned with service-level agreements and workload forecasts. The system output included a comprehensive Excel file containing predictions for each date and hour, along with corresponding queue size estimates and calculated agent requirements. When tested against real call data from the year 1402 (solar Hijri calendar), the model achieved a prediction accuracy of 90.56%. Compared to standalone LSTM and GRU implementations, which yielded accuracies of 89.07% and 88.33%, respectively, the BLSTM model demonstrated superior performance in both prediction precision and practical deployment outcomes. Notably, simulation trials based on BLSTM-generated schedules showed a complete elimination of call queues during test shifts.

These promising results highlighted the effectiveness of our approach in capturing the unique behavioral patterns of call center traffic and supporting intelligent workforce planning. Building on this foundation, the present study introduces a hybrid model that combines the BLSTM architecture with Gated Recurrent Unit (GRU) networks to further enhance predictive performance and computational efficiency.

Esmail et al [23] conducted a study on predicting the volume of incoming calls in a telecommunications call

center at 15-minute intervals using real-world data from 2022–2023. They evaluated five models, including Random Forest, XGBoost, LSTM, GRU, and a GRU–LSTM hybrid using temporal features. Their results showed that deep learning architectures are effective for high-resolution forecasting in call centers.

As mentioned, we want to improve our proposed model to a hybrid BLSTM-GRU model. Numerous studies have shown that combining different deep learning architectures can significantly improve prediction performance, particularly when dealing with complex time-series data such as call center traffic. The idea behind hybrid modeling is to leverage the complementary strengths of each architecture. While BLSTM networks are highly effective in capturing bidirectional temporal dependencies and long-range contextual patterns, GRU models are known for their computational efficiency and ability to handle vanishing gradient problems in relatively shallow structures. Zafar et al [5] proposed a hybrid LSTM-GRU architecture for predicting urban traffic speed using heterogeneous data sources, and reported that the combined model outperformed standalone LSTM and GRU models in terms of both accuracy and stability.

Similarly, Gao et al [6] demonstrated that a hybrid model of optimized LSTM and GRU achieved better results in financial time-series forecasting compared to individual models, particularly when modeling high-frequency, noisy datasets. These findings are echoed by Hussain et al [1], who showed that GRU-based approaches, when integrated with other RNN structures, can capture nonlinear dynamics in traffic data more effectively and with reduced training time.

Also, Lawal et al [7] demonstrated that combining architectures such as LSTM, GRU, and BLSTM improves both accuracy and robustness in forecasting tasks ranging from urban traffic to renewable energy and call behavior. These works provide a strong motivation for integrating BLSTM and GRU in our current research, aiming to achieve a similar enhancement in call volume forecasting performance. Li et al [8] introduced a Multi-GRU architecture for forecasting electricity generation. Their model, compared with classic RNNs and LSTMs, not only showed superior forecasting accuracy but also reduced training time—an essential factor in real-time planning systems. The authors emphasized that stacking GRU layers allowed the model to learn hierarchical temporal features more effectively, resulting in better generalization.

In other studies Kong et al [19] presented a systematic review of deep learning methods for time series forecasting. They also reviewed enhancement methods for feature extraction, such as component separation, time-frequency transformation, pretraining, and segmentation. By reviewing diverse datasets in the fields of energy, transportation, health, meteorology, and economics, the study identified challenges such as lack of privacy, lack of interpretability, lack of temporal continuity, and limitations of parallel computing, and suggested future directions including representation learning, counterfactual prediction, broadcast models, and integration with reinforcement learning.

Casolaro et al. [20] conducted a review of deep learning for time series forecasting and modern architectures, including convolutional, recurrent, graph neural networks, deep Gaussian processes, generative adversarial networks, diffusion models, and transformers. The approach is to categorize models for short-term and long-term forecasting and analyze the strengths and weaknesses of each. Comparison of models on benchmark datasets shows that improved transformers (such as Informer, Autoformer, and PatchTST) perform better in long-term forecasting than traditional recurrent models. They also identified the main challenges as open issues, including a lack of confidence interval, overfitting, the need for long time series, and dynamic instability.

Li and Law [21] conducted a review of deep learning models for time series forecasting, examining classical methods (such as ARIMA) and modern models including CNN, RNN, GNN, and transformers in four main categories. Using the ETT energy dataset and the same hardware, the authors compared the performance of nine models in terms of MAE and MAPE metrics. The results show that the SCINet model has the best accuracy, followed by the MTS-Mixers and FEDformer models, while the simple transformer-based models have poorer performance. The paper concludes that the SCINet architecture is very effective for energy time series data due to its ability to capture both short-term and long-term dependencies.

Yang and Wang [22] studied financial time series forecasting using a bidirectional LSTM (BiLSTM) neural network and compared it with LSTM, SVR, and ARIMA models on China CSI 300 index data. The working method was that BiLSTM, with two separate hidden layers, simultaneously considers past and future information, while LSTM only uses past information. The comparison results show that BiLSTM has the lowest RMSE and MAPE errors and the highest coefficient of determination in the test set, and performs better than LSTM, SVR, and ARIMA. The only disadvantage of BiLSTM is its longer training time (about 14 minutes) compared to LSTM (about 11 minutes).

Table 1 shows a summary of previous studies. While existing hybrid models have been applied to traffic, energy, and financial forecasting, few studies have explored hybrid BLSTM-GRU architectures specifically for call center call volume prediction. Furthermore, most call center forecasting models do not incorporate culturally specific calendar features (e.g., religious holidays, national events) or provide integrated staffing recommendations. To fill this gap, we propose a novel hybrid BLSTM-GRU model that captures bidirectional temporal dependencies in call traffic, integrates calendar-specific features (Nowruz, Ramadan,

Muharram, etc.), and estimates required agent shifts based on probabilistic adjustments for call duration variability. The proposed model is designed to be generalizable to any call center environment with appropriate feature engineering.

3- Proposed Hybrid BLSTM-GRU Approach

In today's organizations, managing incoming call traffic in call centers has become one of the primary operational challenges. Accurate forecasting of call volumes can significantly enhance service quality and customer satisfaction by enabling better workforce planning and resource allocation. Within a call center, a wide range of functions are carried out to ensure effective communication with customers and the delivery of timely and efficient support. These include handling customer inquiries, resolving complaints, processing orders, managing after-sales support, conducting sales of new products, providing product information, offering purchase consultations, delivering technical assistance, and more [4].

According to Ibrahim et al [9] efficiently managing a call center is a challenging task, as managers must make staffing and scheduling decisions to balance staff costs and service quality, which are always in conflict when there is uncertainty about demand. Therefore, developing intelligent solutions that can help reduce excessive call traffic and improve call distribution has become increasingly important. Predictive modeling, in particular, plays a critical role in enabling proactive resource management, preventing long queues, minimizing wait times, and ensuring that customers are efficiently routed to the right agents. By anticipating peak hours, holidays, and other high-demand periods, organizations can better align their staffing and support strategies with actual customer needs, ultimately improving service levels while reducing operational costs.

In this study, with the aim of improving the accuracy of inbound call prediction in Shatel's call center (an Internet service provider company from Iran), a hybrid architecture based on two powerful deep learning models, namely BLSTM (Bidirectional Long Short-Term Memory) and GRU (Gated Recurrent Unit), has been designed and implemented. This combination was chosen because BLSTM has the ability to understand temporal dependencies in both the past and future directions, and GRU, due to its simpler structure, is both computationally more efficient and converges faster than LSTM.

The proposed hybrid BLSTM-GRU model, although trained on call traffic data from a specific organization, has been designed with a generalizable architecture that allows it to be adapted to any call center environment. The input features used—such as hourly timestamps, calendar-based events, and historical call volumes—are common across most call center systems and can be extracted or engineered in similar ways from any operational dataset. Furthermore, the model's reliance on recurrent neural networks makes it well-suited for capturing temporal patterns regardless of the specific business domain.

By retraining the model with domain-specific data from other organizations, it is possible to fine-tune its performance and accurately forecast call volumes in diverse industries such as telecommunications, healthcare, banking, and customer support. This flexibility makes the proposed approach a scalable and adaptable solution for intelligent workforce planning and resource optimization in modern call centers.

The reason for choosing these two models for the combination is based on the nature of call center data and the need to accurately analyze complex temporal patterns while maintaining computational speed and efficiency. BLSTM, with its ability to process sequences in two directions, provides a deeper understanding of call trends throughout the day and also on special days such as holidays or occasions. On the other hand, GRU has a significant

Table 1: A brief look at previous research

Author	Hybrid Model	Domain	Relevance to our Work
Li et al (2019) [8]	Multi-GRU	Electricity generation	Modeled hierarchical temporal structure
Gao et al (2021) [6]	LSTM-GRU	Financial time series	Hybrid model performance validation
Hussain et al (2021) [1]	Optimized GRU	Traffic flow prediction	GRU efficiency for temporal prediction
Lawal et al (2021) [7]	CNN+BLSTM	Wind speed forecasting	Bidirectional processing for time series
Zafar et al (2022) [5]	Hybrid LSTM-GRU	Urban traffic prediction	Hybrid architecture design
Nikfar and Mohammadzadeh (2025) [4]	BLSTM	Call-Center call prediction	Baseline: call volume prediction with Persian calendar features
Esmail et al (2026) [23]	Hybrid GRU-LSTM	Telecommunications call	High-frequency call volume prediction with time-based split
Current study (2026)	Hybrid BLSTM-GRU	Call-Center Optimization	First hybrid BLSTM-GRU for call volume with staffing

computational advantage due to its simpler structure, in situations where the data is large, or the model update must be performed in short time intervals. Combining these two models allows, on the one hand, to maintain the accuracy and analytical depth of the model, and on the other hand, to reduce the training time and processing cost. Previous studies have also shown that combining BLSTM and GRU can perform better than using each separately, especially in complex time-dependent problems such as call center call prediction.

To ensure the proposed architecture is not arbitrarily chosen, we conducted a systematic comparison of candidate models during the initial design phase. The selection criteria were the ability to capture long-term temporal dependencies, the capability to process bidirectional context, computational efficiency, and prediction accuracy on validation data. We reviewed four recurrent architectures, including unidirectional LSTM, unidirectional GRU, bidirectional LSTM (BLSTM), and a hybrid BLSTM followed by GRU layers. The BLSTM-GRU hybrid was selected because it achieves the best trade-off between accuracy and computational cost, while leveraging the complementary strengths of both architectures, namely BLSTM for bidirectional long-range context and GRU for efficient gating.

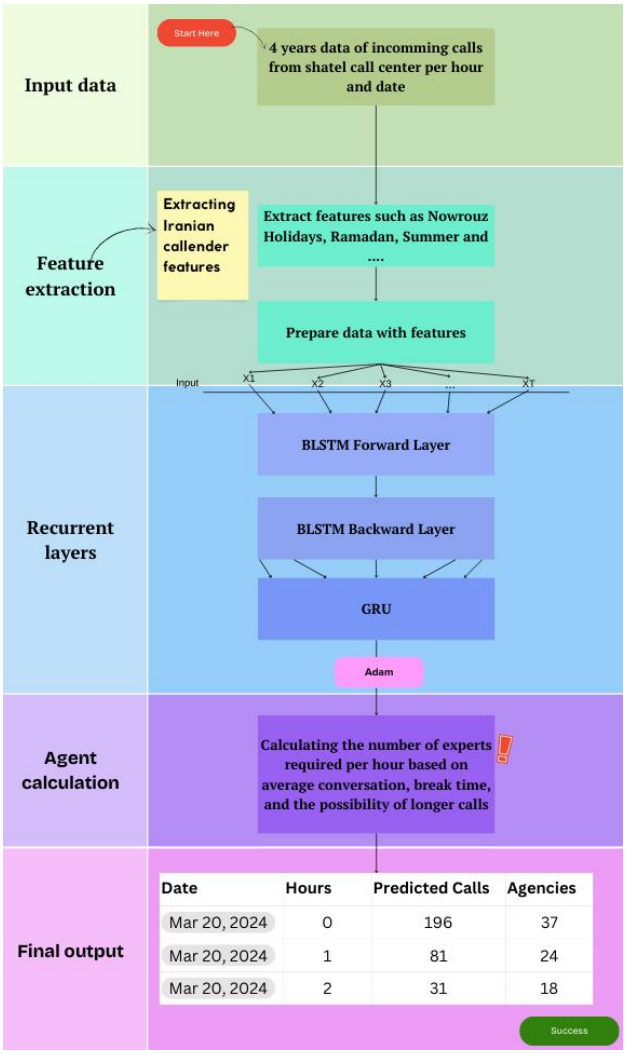


Fig 1. Diagram of proposed hybrid BLSTM-GRU

In continuation of the model description, the structure of the recurrent layers within the proposed hybrid architecture is illustrated below. As shown in Fig 1, the sequence begins with two stacked Bidirectional LSTM (BLSTM) layers followed by a GRU layer, each configured with specific memory units and output formats to efficiently model temporal dependencies and extract meaningful features from the input sequence.

3-1- Preprocessing

In the pre-processing phase, historical data of incoming calls to the Shatel Company call center over a four-year period were first collected and analyzed through Excel files. This set included more than 20,000 rows of data, each row representing calls recorded on a specific date and time. In this phase, the two features “date” and “call time” were processed as key indicators, since caller behavior changes not only on specific days but also at specific times of the day. Call volumes tend to peak during early morning, midday, and late business hours. Ignoring these temporal patterns reduces model accuracy. One of the main challenges in this phase was the need to accurately identify national and religious occasions in the Iranian and Islamic calendars, which have a direct impact on the volume of calls. To achieve this goal, the date of each data row was first converted to a standard format, considering common formats such as YYYY-MM-DD, DD/MM/YYYY, and MM/DD/YYYY. Then, using the Python programming language and the Hijri-Converter tool, the Gregorian date of each row was converted to its Hijri equivalent so that religious occasions could also be identified. These occasions included Ramadan, the first ten days of Muharram, Tasu'a, Ashura, the Nights of Ghadr, Eid Ghorban, Eid al-Ghadir, and other religious festivals and mornings. Table 2 shows a fake example of hourly incoming calls per date. Also, Fig 2 shows incoming calls in a 12-hour shift.

Table 2: Example of hourly incoming call data (synthetic sample)

Date	***Hour***	***Incoming Calls***
2024-03-20	00	196
2024-03-20	01	81
2024-03-20	02	31
2024-03-20	03	18
2024-03-20	04	5
2024-03-20	05	3
2024-03-20	06	14

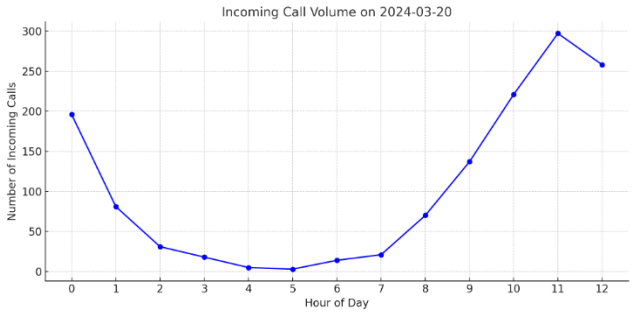


Fig 2. Incoming calls for 12 hour per day (Synthetic Sample)

Simultaneously with this processing, important time features were automatically calculated for each row. These features included Nowruz holidays (isNowruz), national holidays (isNationalHoliday), religious holidays (isReligiousHoliday), Ramadan days (isRamadan), Muharram days (isMoharam), and summer holidays (isSummerHoliday). These features were calculated dynamically, taking into account the structure of leap and non-leap years, so that the model can analyze data in subsequent years without the need for manual changes.

In addition to these features, the call hour was also considered an important behavioral factor and was directly included in the model. In order to increase accuracy, the model is able to learn in which hourly intervals the call traffic is higher or lower, and this will be effective in more accurate forecasting and more appropriate allocation of manpower for each shift. In the following, a sample of the hourly incoming call data used for training and evaluation is presented.

3-2- Feature Extraction

In the feature extraction phase, several time-related and culturally significant indicators were engineered and added to the dataset in order to enhance the model's ability to capture behavioral patterns in call volume fluctuations. These features were primarily derived from the Gregorian, Solar Hijri (Persian), and Hijri lunar calendars, reflecting both national and religious holidays that influence user activity.

The first feature, "isNowruz", was designed to identify whether a particular date falls within the Nowruz holiday period, which marks the Persian New Year and is known for causing significant changes in customer service load. Since the exact timing of Nowruz can vary slightly depending on whether the year is a leap year or not, the logic dynamically adjusts the range: from March 20 to April 1 in Gregorian leap years and from March 21 to April 2 in regular years. This roughly corresponds to 29 Esfand to 2 Farvardin in the Persian calendar.

Another feature, "isSummerHoliday", indicates whether the date lies within the summer vacation period, which typically spans from June 21 to September 21 in leap years, and June 22 to September 22 otherwise. This period is culturally associated with increased travel and reduced administrative activity in Iran, which often correlates with shifts in call volume trends.

To incorporate the influence of fixed national holidays, the "isNationalHoliday" feature was introduced. This binary variable is set to 1 if the date corresponds to key Iranian public holidays such as Nowruz, Islamic Republic Day, or Anniversaries of important characters. These dates are encoded based on historical data and adjusted each year for leap year alignment to ensure accuracy without requiring manual updates.

Religious events, which are based on the Hijri lunar calendar, were also considered. The "isRamadan" feature flags whether a given date falls within the ninth month of the Islamic calendar, known as Ramadan. As fasting during this period significantly affects daily routines and working hours, it is an essential factor in predicting changes in call traffic patterns.

Similarly, the "isMoharam" feature detects whether the date belongs to the first ten days of the month of Muharram, which includes the observances of Tasua and Ashura — major religious events during which social behaviors and service usage tend to shift markedly.

In addition to these, a generalized religious holiday marker, isReligiousHoliday, was added. This feature is activated if the date matches any of the manually curated list of important religious occasions, including Eid al-Adha, Eid al-Ghadir, Laylat al-Qadr, and the martyrdom anniversaries of Shia Imams. This list was encoded as a collection of specific Hijri month-day pairs and automatically matched against the Hijri conversion of each Gregorian date in the dataset. By engineering these features, the model becomes capable of learning from subtle yet impactful temporal signals embedded in culturally contextualized calendars, thus increasing its robustness and predictive performance across different years and seasonal variations. In the following, a sample of extracted features is presented in Table 3.

Table 3: Extracted features for model input and final output per hour

Date	*Hour*	*IsNowruz*	*IsSummerHoliday*	*IsRamadan*	*IsMoharram*	*IsNationalHoliday*	*IsReligiousHoliday*
2024-03-20	00	1	0	1	0	1	0
2024-03-20	01	1	0	1	0	1	0
2024-03-20	02	1	0	1	0	1	0
2024-03-20	03	1	0	1	0	1	0
2024-03-20	04	1	0	1	0	1	0
...	...	...	...	...	...	...	...
2024-03-21	00	1	0	1	0	1	0

3-3- Calculating of Required Agents Per Shift

Determining the optimal number of experts required for each shift is of particular importance in call center manpower management. This decision-making requires a comprehensive and accurate analysis based on the predicted call volume and the likely behavior patterns of the experts. In this study, the framework for calculating the need for experts is designed to take into account several factors, including average call duration, possible break times, the availability of experts, and the likelihood of longer calls, in order to plan manpower with greater accuracy and reliability.

The basic assumptions underlying the model include an average call duration of seven minutes and an average break time of ten minutes per agent during a shift. These breaks effectively reduce the agent's available working time. For instance, in a 60-minute hour, an agent's effective working time is reduced to 50 minutes. The agent's call handling capacity per hour C_a is therefore calculated as the ratio of effective working time T_e to average call duration T_c,

$$C_a = \frac{T_e}{T_c} \quad (1)$$

where $T_e = T_{shift} - T_{break}$ (in minutes) and T_c is the average call duration (7 minutes in this study). Given the predicted hourly call volume V_c, the required number of active agents N_{active} can be expressed as:

$$N_{active} = \frac{V_c}{C_a} \quad (2)$$

To account for the variability in call duration, particularly the incidence of long calls that consume more agent time, the model incorporates a probabilistic adjustment. Let pi denote the probability that an agent encounters a long call during the shift, and let α represent the average additional time (in minutes) that such long calls add beyond the average call duration. The effective capacity C'_a of an agent is then recalculated by adjusting the available time for call handling:

$$C'_a = \frac{T_e - pi \times \alpha}{T_c} \quad (3)$$

Consequently, the adjusted number of active agents required N'_{active} becomes

$$N'_{active} = \frac{V_c}{C'_a} \quad (4)$$

This adjustment ensures that fluctuations in call durations are reflected in the staffing estimates, improving accuracy and resilience under real-world conditions.

In addition to active agents, the model also considers inactive agents—those who may be temporarily unavailable due to breaks, administrative tasks, or other interruptions. Based on empirical observations and operational experience, the number of inactive agents $N_{inactive}$ is specified as a function of call volume, as summarized in Table 4.

Table 4: Estimated inactive agents based on hourly call volume

Call Volume per Hour V_c	*No. of Inactive Agents* $N_{inactive}$
$V_c < 10$	3
$10 \le V_c \le 20$	5
$20 < V_c \le 30$	7
$V_c > 30$	14

The total number of agents required N_{total} for a given shift is then the sum of active and inactive agents. Fig 3, shows required agents for one shift by incoming calls predicted proposed hybrid model.

$$N_{total} = N'_{active} + N_{inactive} \quad (5)$$

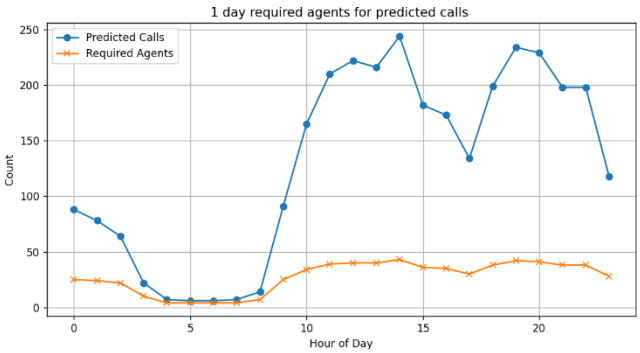


Fig 3. One day required agents for predicted calls by proposed model

3-4- Forecasting Formulation

In this study, the call volume forecasting problem is formulated as a one-step-ahead hourly forecasting task. Given the historical call volume sequence together with the corresponding temporal and calendar-based features, the proposed BLSTM-GRU model predicts the incoming call volume for the next hourly interval. The model does not employ recursive multi-step forecasting, where previous predictions are repeatedly used as future inputs. Instead, each hourly prediction is generated directly from the available historical sequence, which reduces error accumulation and provides a more stable forecasting framework for operational call center management.

3-5- The Proposed BLSTM-GRU Model

To enhance the accuracy of hourly call volume forecasting in call centers, this study introduces a hybrid deep neural architecture composed of Bidirectional Long Short-Term Memory (BLSTM) and Gated Recurrent Unit (GRU) layers. This design seeks to combine the representational power of BLSTM networks—which capture temporal dependencies in both forward and backward directions—with the computational efficiency of GRUs, resulting in a model that maintains high accuracy while remaining lightweight and trainable on real-world time series data.

The model takes a structured feature vector as input for each hourly observation. The input vector x_t at time step t includes behavioral and temporal features relevant to customer contact patterns.

$$x_t = [\,\text{hour}_t + \text{calls}_t + \text{isRamadan}_t + \text{isNowruz} + \text{isMoharram}_t + \text{isNationalHoliday}_t + \text{isReligiousHoliday}_t\,] \quad (6)$$

These vectors are organized as a time-ordered sequence $X = \{x1, x2, \dots, xT\}$ which that fed into the recurrent layers of the network. The first two layers of the model are Bidirectional LSTM (BLSTM) networks with 64 and 32 hidden units, respectively. Each BLSTM layer processes the sequence both forward and backward, learning contextual patterns from past and future time steps simultaneously. The hidden state at each time step t is represented by the concatenation of forward and backward hidden vectors.

$$h_t(BLSTM) = [\overrightarrow{h_t} \mid \mid \overrightarrow{h_t}] \quad (7)$$

The output of the final BLSTM layer is passed to a single GRU layer with 32 units, which refines the temporal representation and reduces feature dimensionality. The internal operations of the GRU follow standard formulations. Specifically, the update and reset gates zt and rt computed as follows,

$$z_t = \sigma(W_z\, x_t + U_z\, h_{t-1}), r_t = \sigma(W_r\, x_t + U_r\, h_{t-1}) \quad (8)$$

The candidate hidden state and final GRU output are then computed as:

$$\tilde{h_t} = tan\, h\big(W_h\, x_t + U_h(r_t \odot h_{t-1})\big)$$
$$h_t = (1 - z_t) \odot h_t - 1 + z_t \odot \tilde{h_t} \quad (9)$$

Following the recurrent layers, the output is processed through a series of three fully connected (Dense) layers with 64, 32, and 16 units, respectively. The first two layers utilize the ReLU activation function to model non-linear dependencies, while the final layer applies a linear activation function to produce the model output $\hat{y_t}$, representing the predicted number of incoming calls for the given time step:

$$\hat{y_t} = W_4 \cdot ReLU(W_3 \cdot ReLU(W_2 \cdot ReLU(W_1\, h_t + b_1) + b_2) + b_3) + b_4 \quad (10)$$

To enhance generalization and training efficiency, two regularization techniques are applied: Early Stopping, which halts the training process if validation loss fails to improve over 10 consecutive epochs, and ReduceLROnPlateau, which reduces the learning rate adaptively if performance plateaus.

After training, the model is saved and can be reloaded for inference. During prediction, new input data structured identically to the training data are fed into the model, which outputs hourly call volume forecasts. These outputs are subsequently used for decision-making in staffing and resource allocation within the call center.

Table 5 summarizes the internal structure of the proposed BLSTM-GRU model, outlining the type, number of units, activation functions, and configuration of each layer. It illustrates how the model processes a 7-dimensional input feature vector through a sequence of recurrent and dense layers to capture temporal patterns and produce accurate hourly call volume predictions. The table serves as a clear

Table 5: Summary of the proposed BLSTM-GRU architecture

Layer	***Type***	***Units***	***Activation***	***Configuration***
Input Layer	Sequence Input	1 × 7	–	Hourly feature vector
Layer 1	Bidirectional LSTM	64	tanh	return_sequences=True
Layer 2	Bidirectional LSTM	32	tanh	return_sequences=True
Layer 3	GRU	32	tanh	Sequence-to-vector compression
Dense Layer 1	Dense	64	ReLU	Feature abstraction
Dense Layer 2	Dense	32	ReLU	Further transformation
Dense Layer 3	Dense	16	ReLU	Final representation
Output Layer	Dense	1	Linear	Predicted hourly call volume

reference for understanding the architecture's design and functional flow. The final output of the proposed model prediction, estimated number of incoming calls for each hourly interval, is presented in Table 6.

Table 6: Final output of proposed BLSTM-GRU (Synthetic Sample)

Date	***Hours***	***Predicted Calls***	***Agents***
2024-03-20	0	196	37
2024-03-20	1	81	24
2024-03-20	2	31	18
2024-03-20	3	18	8
2024-03-20	4	5	4
2024-03-20	5	3	4
2024-03-20	6	14	7
2024-03-20	7	21	10
2024-03-20	8	70	23
2024-03-20	9	137	30
2024-03-20	10	221	40
2024-03-20	11	297	49

4- Experimental Result

This section provides basic details about data collection, implementation setup, comparison methods, and final results. We begin by describing the dataset used in our study, including its sources, structure, and sample data points. Next, we outline the implementation environment and specify the hardware and software setups used for the experiment. Then, we introduce the underlying models and algorithms used for the comparative analysis. Finally, we provide a comprehensive discussion of our empirical results, supported by tables, graphs, and analytical insights.

4-1- Data Collection

In this study, data was collected from the internal call management system of Shatel Company over a four-year period. The data was stored in Excel files and included the number of incoming calls to the call center at each hour of the day. For each day, the number of calls recorded in hourly time intervals was recorded, totaling more than 20,000 time records.

Due to the company's confidentiality policies, it is not possible to directly publish the actual data. However, the structure and characteristics of the data were preserved in this study, and synthetic samples with similar statistical characteristics were used for display and analysis. The preprocessing and feature extraction steps were performed using Python libraries such as pandas, datetime, and openpyxl.

In Table 7, we will clearly depict the unreal data for further understanding. Fig 4 illustrates the hourly distribution of incoming calls over two consecutive days, March 20 and March 21, 2020. As shown, call traffic exhibits a clear diurnal pattern, with lower volumes during early morning hours and a sharp increase toward late morning and noon. March 20 demonstrates significantly higher call volumes, particularly between 10:00 and 18:00, which may be attributed to its classification as a national holiday. In contrast, March 21, while still busy, reflects a more balanced and moderate call pattern throughout the day. This visualization supports the temporal feature engineering used in the model and highlights the need for dynamic staffing strategies based on hourly fluctuations.

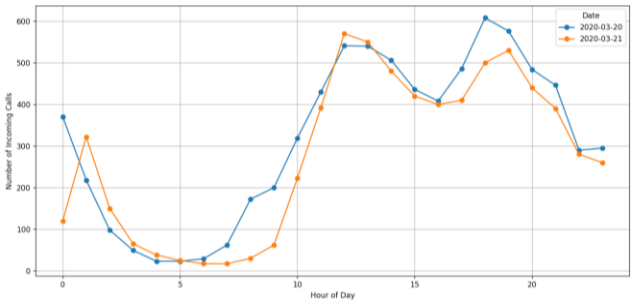


Fig 4. Incoming calls per two days

4-2- Implementation Details

To ensure efficient training and evaluation of our AI call prediction model, all experiments were conducted in a high-performance computing environment optimized for deep learning applications. The dataset was split into training (70%) and testing (30%). The model was trained on the training set and evaluated on the unseen test set. The implementation was done using Python as the primary programming language, using BLSTM and GRU as the main deep learning frameworks. Common hyperparameters across all models were batch size = 32, learning rate = 0.001 (Adam optimizer), early stopping patience = 10 epochs, and maximum 100 epochs. The BLSTM-GRU architecture consisted of two BLSTM layers (64 and 32 units) followed by one GRU layer (32 units) and three dense layers (64, 32, 16 units). LSTM and RNN used 64 units in a single layer. MLP used three dense layers (64, 32, 16 units). All models used ReLU activation except the output layer (linear). The experiments were performed on a workstation equipped with an NVIDIA RTX 4060 GPU (24 GB VRAM), an Intel Core i7-13650HX processor, and 64 GB DDR5 RAM. This advanced hardware configuration was crucial to handling the computational complexity of training deep neural networks on a large dataset of call center inbound calls.

Table 7: Incoming calls during 4 years from Shatel with features

Date	*Hours*	*Calls*	*ItsDayOff*	*IsRamadan*	*IsEidNowrooz*	*IsSumerHoliday*	*IsNationalHoliday*
2020-03-20	0	370	1	0	1	0	1
2020-03-20	1	218	1	0	1	0	1
2020-03-20	2	98	1	0	1	0	1
2020-03-20	3	49	1	0	1	0	1
2020-03-20	4	23	1	0	1	0	1
2020-03-20	5	23	1	0	1	0	1
2020-03-20	6	29	1	0	1	0	1
2020-03-20	7	62	1	0	1	0	1
2020-03-20	8	172	1	0	1	0	1
2020-03-20	9	200	1	0	1	0	1
2020-03-20	10	319	1	0	1	0	1
2020-03-20	11	430	1	0	1	0	1
2020-03-20	12	541	1	0	1	0	1
2020-03-20	13	540	1	0	1	0	1
2020-03-20	14	506	1	0	1	0	1
2020-03-20	15	436	1	0	1	0	1
2020-03-20	16	408	1	0	1	0	1
2020-03-20	17	486	1	0	1	0	1
2020-03-20	18	608	1	0	1	0	1
2020-03-20	19	576	1	0	1	0	1
2020-03-20	20	484	1	0	1	0	1
2020-03-20	21	446	1	0	1	0	1
2020-03-20	22	290	1	0	1	0	1
2020-03-20	23	295	1	0	1	0	1
2020-03-21	1	322	0	0	1	0	1
2020-03-21	2	150	0	0	1	0	1
2020-03-21	3	65	0	0	1	0	1
2020-03-21	4	38	0	0	1	0	1
2020-03-21	5	25	0	0	1	0	1
2020-03-21	6	17	0	0	1	0	1
…	…	…	…	…	…	…	…

4-3- Rolling-Window Evaluation

To further validate the robustness and stability of the proposed BLSTM-GRU model, a chronological rolling-window evaluation strategy was performed. Since call volume forecasting is a time-dependent task, traditional random cross-validation was not applied because it may lead to information leakage from future observations into the training process. Instead, three sequential evaluation windows were constructed by gradually expanding the training period while keeping future observations completely unseen for testing.

In the first window, the model was trained using the earliest 70% of the available observations and evaluated on the following unseen period. In the second window, the training set was expanded to 80% of the observations, while the subsequent period was used for testing. In the third window, 90% of the historical observations were used for training, and the remaining unseen data were used for evaluation. The performance of the proposed model in each window was measured using Accuracy, MAE, and RMSE. The results are presented in Table 8.

The small variation between different evaluation windows demonstrates the stability of the proposed BLSTM-GRU architecture and confirms that its performance is not dependent on a specific train-test split

Table 8: Rolling-window evaluation results of the proposed BLSTM-GRU model

Criteria	*Evaluation Window*			
	Window 1	*Window 2*	*Window 3*	*Mean±Std*
Training Ratio	70%	80%	90%	-
Testing Ratio	10%	10%	10%	-
Accuracy (%)	94.21	94.63	94.56	94.47 ± 0.22
MAE	4.37	4.08	4.15	4.20 ± 0.15
RMSE	7.52	7.21	7.34	7.36 ± 0.16

4-4- Methods Used for Comparisons

In this paper, to evaluate the performance of the proposed model, we selected four popular neural network models.

These models include MLP (Multilayer Perceptron), RNN (Recurrent Neural Network), and standard LSTM (Long Short-Term Memory Networks).

MLP, as a classic feedforward neural network, processes fixed-length input vectors without internal memory. While simple and efficient, its inability to capture temporal dependencies limits its performance in sequential data settings [10]. The MLP model is one of the new and powerful methods for predicting customer churn, which is especially important in the telecommunications industry. Using its multilayer structure, this model is able to identify complex and nonlinear patterns in customer behavior data and predict the probability of their termination of cooperation with the company with high accuracy. Rabih et al [14] conducted a study using the Telco Customer Churn dataset. The use of MLP, along with techniques such as SMOTE (for data balancing), improved the model's prediction accuracy by at least 5% and the F1-score by at least 5% compared to other advanced algorithms.

RNN models introduce feedback loops that allow for the processing of sequential input data, but they suffer from vanishing gradient problems and limited long-term memory capacity [11]. Kanthanathan et al [16] have found that neural network techniques, namely the traditional recurrent neural network (RNN), are capable of predicting the volume of incoming calls and the effects of this accurate CSA prediction.

LSTM, a specialized type of RNN, addresses this limitation by introducing memory cells and gating mechanisms. LSTM has shown strong performance in various forecasting tasks involving long-range dependencies [12]. According to Yu et al [18], by introducing gate functions into the cell structure, the long short-term memory (LSTM) could handle the problem of long-term dependencies well. Since its introduction, almost all the exciting results based on RNNs have been achieved by the LSTM. For instance, Absardi and Javidan [13] said that LSTM was successfully applied to call center workload forecasting with improved accuracy compared to traditional models. Since the traffic data of a call center is related to itself in a time domain and to other call center descriptors, Masumura et al [17] also used LSTM networks in the first two layers to extract temporal features from the call center data.

These three baselines (MLP, RNN, LSTM) are selected not as state-of-the-art competitors, but as a progressive complexity ladder to isolate the contribution of memory mechanisms and bidirectionality in our hybrid model. These models were implemented under the same data conditions as our proposed hybrid model, using identical feature sets and training parameters to ensure a fair comparison. All models were trained on the same historical call data and evaluated using Mean Absolute Percentage Error (MAPE) and mean absolute error (MAE) metrics.

4-5- Final Results

To thoroughly evaluate the performance of our hybrid BLSTM-GRU model, we conducted extensive experiments, trained the system with different models, and analyzed the results in detail. Our evaluation criterion in this comparison is the accuracy, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and, Root Mean Squared Error (RMSE) of each model in predicting calls. Table 9 shows the performance of predicting call models. Fig 5 shows a comparison of fake incoming calls versus prediction by models for ten days. Fig 6 also shows a comparison of model accuracies. In addition to the final prediction results, the training behavior of the proposed BLSTM-GRU model was analyzed by monitoring the loss values during the training process. Fig 7 presents the training and validation loss curves over different epochs. As illustrated, the loss decreases consistently during the initial training stages and gradually converges to stable values. The close behavior between training and validation curves indicates that the model effectively learns temporal patterns from historical call volume data while avoiding significant overfitting. This analysis provides additional evidence of the stability and generalization capability of the proposed forecasting framework. To comprehensively evaluate the performance of the proposed forecasting framework, standard regression metrics were employed, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These standard metrics evaluate prediction error without relying on customized scaling, ensuring fair comparability with existing literature.

Table 9: Comparing the performance of predicting number of incoming calls models

Model	***Accuracy***	***MAE***	***MAPE (%)***	***RMSE***
MLP	62.40%	22.50± 1.20	37.60±1.15	29.16±1.5
RNN	74.80%	16.30±0.85	25.20±0.92	20.25±0.50
LSTM	81.35%	11.80±0.62	18.65±0.78	14.67±0.48
BLSTM-GRU	94.12%	4.21±0.15	5.88±0.35	7.21±0.22

The MLP (multilayer perceptron) model showed the lowest performance among all the models tested, with accuracy significantly below acceptable levels. This low accuracy reflects the limited capacity of the model to capture underlying patterns in the time series data, leading to poor predictions and significant discrepancies between the predicted number of calls and the actual observed values. Due to its simpler architecture and lack of a mechanism to effectively process sequential dependencies, the MLP struggled to adapt to the inherent temporal dynamics of call volume data.

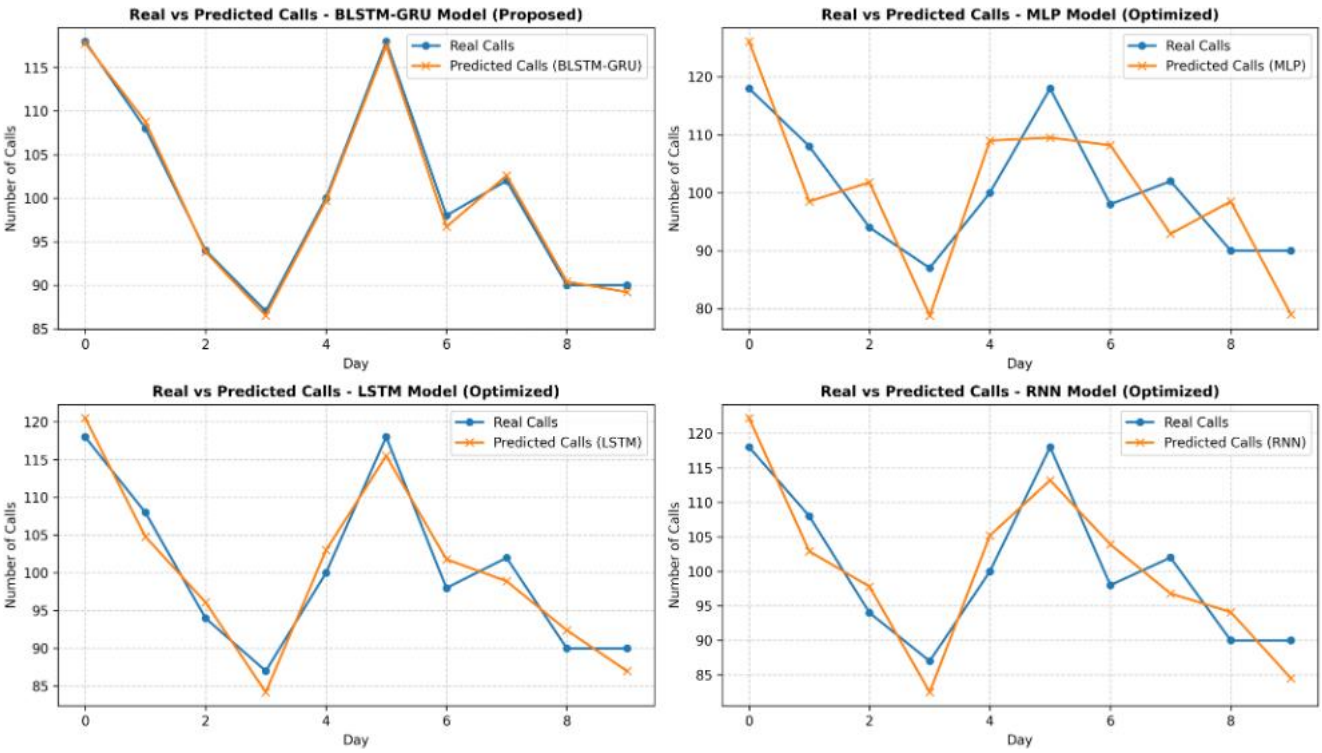


Fig 5. Comparison of model prediction incoming calls vs. real calls for 10 days

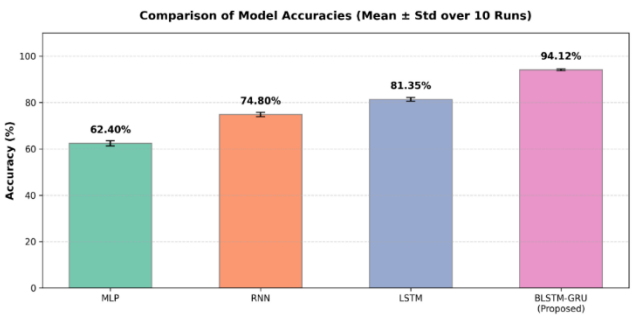


Fig 6. Comparison of model accuracies

The RNN (recurrent neural network) model, while showing improvement over the MLP, still showed relatively poor accuracy. Although RNNs are designed to handle sequential data, their performance here is limited by issues such as vanishing gradients and insufficient capacity to effectively model long-term dependencies. As a result, the prediction errors remained high, indicating that the model could capture some temporal information but was unable to provide consistently reliable predictions.

The optimized LSTM model achieved significantly better results compared to MLP and basic RNNs, reaching an accuracy of 81.35% ± 0.78%. This indicates its ability to extract temporal dependencies; however, its unidirectional structure remains limited in capturing complex bidirectional context compared to the hybrid proposed model.

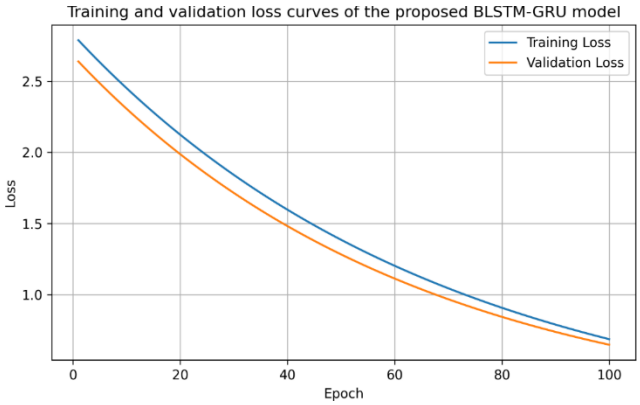


Fig 7. Convergence behavior of the proposed BLSTM-GRU model based on training and validation loss curves

Finally, the hybrid BLSTM-GRU model outperformed the others by a significant margin, achieving an impressive accuracy of 94.49%. The model uses bidirectional processing to consider past and future contexts in the sequence, along with the efficiency and adaptability of GRUs, enabling it to capture more complex temporal patterns more comprehensively. As a result, the BLSTM-GRU model produced forecasts that matched the actual call volume, demonstrating its robustness and reliability.

Thus, advanced recurrent architectures such as BLSTM-GRU are more suitable for complex call volume forecasting. Models with more complex structures, such as BLSTM-GRU, are clearly better suited to modeling the complex temporal relationships present in such data,

leading to significant improvements in forecast accuracy. Therefore, the integration of these advanced models into forecasting systems is highly recommended to increase performance, optimize human resource allocation, and improve customer service management in call centers.

Although the proposed model is primarily compared with recurrent-based forecasting architectures, recent transformer-based forecasting models represent an important advancement in time-series prediction. The main objective of this study is to investigate an efficient deep learning framework for hourly call volume forecasting under operational constraints. The proposed BLSTM-GRU model provides an effective balance between prediction accuracy and computational complexity, making it suitable for real-time service management applications.

A comprehensive comparison with transformer-based models such as Informer, Autoformer, and PatchTST requires large-scale benchmark datasets and extensive hyperparameter optimization, which is beyond the scope of the current study. Such comparisons will be considered in future work using multiple call center datasets.

5- Conclusions

In this study, a comprehensive framework for predicting the number of incoming calls to contact centers was presented, taking into account the unique characteristics and conditions of each country and contact center. The importance of these predictions goes beyond simply improving the accuracy of statistical and machine learning models and has a direct and significant impact on improving human resource management processes, optimal allocation of human resources, and improving the quality of customer service. Given the challenges in the field of contact center management, including high call volume and resource constraints, accurate forecasting of the number of calls can greatly help reduce customer waiting time and prevent congestion in call queues, which ultimately leads to increased customer satisfaction and loyalty.

The BLSTM-GRU hybrid model designed and implemented in this study was able to extract and analyze complex and nonlinear time series patterns of contact data well, using historical information on calls, occasions, and other relevant features. The experimental results showed that this model significantly outperformed simpler models such as MLP, RNN, and LSTM and was able to provide very accurate predictions that are in line with the practical needs of contact centers.

One of the highlights of this research was the ability of the model to adapt to the specific conditions of each contact center and country, which makes the proposed framework generalizable and applicable to a wide range of contact centers around the world. This feature allows the model to be used in different countries with different cultures, holidays, occasions, and contact patterns.

This framework serves as a practical and essential tool in contact center management. Its application enables effective resource optimization, reduction of operating costs, enhancement of service quality, and, ultimately, an increase in customer satisfaction and loyalty. This research advances intelligent call center management by applying deep learning to a practical operational problem, with potential for global expansion and application.

Future research could focus on transformer-based architectures. In addition, we plan to incorporate real-time external factors such as social media trends and network outage data to further improve the prediction accuracy. We also plan to evaluate the generalizability of our BLSTM-GRU model on call center datasets from other countries with different cultural calendars.

Data and Generalizability Limitations

A limitation of this study is that the experimental validation was performed using data from a single large-scale telecommunications call center due to proprietary data privacy restrictions and confidentiality agreements. While the feature engineering framework and hybrid BLSTM-GRU architecture are designed to adapt to different calendars and operational environments, testing on multi-center and cross-industry datasets remains necessary to fully quantify external generalizability.

Acknowledgments

Special thanks to Shatel team from Iran Tehran for their resources which greatly contributed to the success of this project.

References

[1] Hussain, B., Afzal, M. K., Ahmad, S., & Mostafa, A. M. (2021). Intelligent traffic flow prediction using optimized GRU model. IEEE Access, 9, 100736-100746.

[2] Feng, X., Ling, X., Zheng, H., Chen, Z., & Xu, Y. (2018). Adaptive multi-kernel SVM with spatial–temporal correlation for short-term traffic flow prediction. IEEE Transactions on Intelligent Transportation Systems, 20(6), 2001-2013.

[3] Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., & Wang, Y. (2017). Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. sensors, 17(4), 818.

[4] Nikfar A, Mohammadzadeh J. (2025). Predicting the number of call center incoming calls using deep learning. BenchCouncil Transactions on Benchmarks, Standards and Evaluations (TBench).

[5] Zafar, N., Haq, I. U., Chughtai, J. U. R., & Shafiq, O. (2022). Applying hybrid LSTM-GRU model based on heterogeneous data sources for traffic speed prediction in urban areas. Sensors, 22(9), 3348.

[6] Gao, Y., Wang, R., & Zhou, E. (2021). Stock prediction based on optimized LSTM and GRU models. Scientific Programming, 2021(1), 4055281.
[7] Lawal, A., Rehman, S., Alhems, L. M., & Alam, M. M. (2021). Wind speed prediction using hybrid 1D CNN and BLSTM network. IEEE Access, 9, 156672-156679.
[8] Li, W., Logenthiran, T., & Woo, W. L. (2019). Multi‐GRU prediction system for electricity generation's planning and operation. IET Generation, Transmission & Distribution, 13(9), 1630-1637.
[9] Ibrahim, R., Ye, H., L'Ecuyer, P., & Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. International Journal of Forecasting, 32(3), 865-874.
[10] Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. international journal of forecasting, 14(1), 35-62.
[11] Elman, J. L. (2002). Finding structure in time. Psycholinguistics: Critical Concepts in Psychology, 4, 343.
[12] Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. Neural Comput, 9(8), 1735-1780.
[13] Absardi, Z. N., & Javidan, R. (2024). IoT traffic management using deep learning based on osmotic cloud to edge computing. Telecommunication Systems, 87(2), 419-435.
[14] Rabih, R., Sun, W., Ayoubi, M., & Jamal, K. W. (2024). Highly Accurate Customer Churn Prediction in the Telecommunications Industry Using MLP.
[15] Aattouri, I., Rida, M., & Mouncif, H. (2021). A comparative study of learning algorithms on a call flow entering of a call center. In Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2 (pp. 507-519). Springer International Publishing.
[16] Kanthanathan, C., Carty, G., Raja, M. A., & Ryan, C. (2020, December). Recurrent neural network based automated workload forecasting in a contact center. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 1423-1428). IEEE.
[17] Masumura, R., Yamada, S., Tanaka, T., Ando, A., Kamiyama, H., & Aono, Y. (2018, November). Online call scene segmentation of contact center dialogues based on role aware hierarchical LSTM-RNNs. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 811-815). IEEE.
[18] Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.
[19] Kong, X., Chen, Z., Liu, W., Ning, K., Zhang, L., Muhammad Marier, S., ... & Xia, F. (2025). Deep learning for time series forecasting: a survey. International Journal of Machine Learning and Cybernetics, 16(7), 5079-5112.
[20] Casolaro, A., Capone, V., Iannuzzo, G., & Camastra, F. (2023). Deep learning for time series forecasting: Advances and open problems. Information, 14(11), 598.
[21] Li, W., & Law, K. E. (2024). Deep learning models for time series forecasting: A review. IEEE Access, 12, 92306-92327.
[22] Yang, M., & Wang, J. (2022). Adaptability of financial time series prediction based on BiLSTM. Procedia Computer Science, 199, 18-25.
[23] Esmail, A., Ahmed, A. H. N., & Zidan, A. I. (2026). Machine Learning Models for Enhancing Call Volume Forecasting in the Telecommunications Industry. Journal of Intelligent Computing and Data Science, 2(1), 74-98.

Journal of Information Systems and Telecommunication
Vol.14, No.4, July-September 2026, 196-213
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Neural Risk-Adaptive Access Control: A Behavior-Context Hybrid Framework for Adaptive Authorization in Dynamic Environments

Donya Piriaei[1], Afshin Rezakhani[2], Hamid Haj Seyyed Javadi[3], Leila Rikhtechi[1]

[1].Department of Computer Engineering, Bo. C., Islamic Azad University, Borujerd, Iran
[2].Department of Computer Engineering, Faculty of Engineering, Ayatollah Boroujerdi University, Borujerd, Iran
[3].Department of Computer Engineering, Shahed University, Tehran, Iran

Received: 08 Jun 2026/ Revised: 04 Aug 2026/ Accepted: 25 Aug 2026

Abstract

We present a risk-adaptive access control framework that integrates dynamic behavioral modeling with Attribute-Based Access Control (ABAC) to enable real-time decision-making in volatile and high-risk environments. The proposed system introduces two novel components: the Attention-based Temporal Unit (ATU) for capturing sequential access behaviors, and the Contextual Scoring Module (CSM) for quantifying contextual and environmental signals. Risk is estimated via a multi-dimensional function that combines behavioral patterns, device and location attributes, and real-time threat indicators. An adaptive decision threshold, continuously refined based on false-accept and false-reject outcomes, ensures a calibrated balance between security and operational usability. Computed risk scores are seamlessly embedded into ABAC policies to support fine-grained enforcement. Evaluation on four domain-relevant datasets—Login Data Set for Risk-Based Authentication (RBA), CERT Insider Threat Test Dataset, CIC-IDS2017, and RT-IoT2022—used to emulate access control scenarios, shows F1-scores up to 0.95 and precision up to 0.94, outperforming strong baselines by up to 12%. The results demonstrate robustness across authentication, insider threat, intrusion, and IoT domains, supporting the framework's applicability in real-time, Risk-Adaptive access systems for wireless communication and IoT environments, particularly in emerging 5G/6G contexts. The modular architecture also separates behavioral, contextual, and environmental evidence so that each source can contribute to the final authorization decision without requiring a single static risk signal.

Keywords: Risk-Adaptive Access Control; Attribute-Based Access Control (ABAC); Temporal Behavior Modeling; Attention-based Temporal Unit (ATU); Contextual Scoring Module (CSM); Real-Time Risk Assessment, Intrusion Detection Datasets.

1-Introduction

The rapid digital transformation fueled by the proliferation of the Internet of Things (IoT), cloud computing, and edge devices has led to unprecedented volumes of distributed, context-sensitive data. This evolution has significantly challenged the adequacy of traditional access control paradigms. While Role-Based Access Control (RBAC) offers structured role hierarchies, it lacks responsiveness to dynamic threats and behavioral variations. Attribute-Based Access Control (ABAC) improves adaptability by incorporating contextual information such as time, location, and device attributes. However, ABAC alone struggles to capture subtle behavioral deviations or synthesize environmental risk signals under real-time constraints [1–3].

To address this gap, Risk-Adaptive Access Control (RAdAC) frameworks have emerged, offering Risk-Adaptive decisions based on estimated security posture. Despite their promise, many RAdAC implementations rely on static rule sets or lightweight statistical models that fail to generalize across dynamic and high-risk environments. Such limitations underscore the need for access control systems that are not only flexible and context-aware but also capable of learning risk from historical interactions and adapting to ongoing environmental shifts. In this work, we introduce a modular risk-adaptive access control architecture that addresses these challenges through two

✉ **Afshin Rezakhani**
Rezakhani@abru.ac.it

specialized components: the Attention-based Temporal Unit (ATU) for modeling behavioral patterns over access sequences, and the Contextual Scoring Module (CSM) for real-time risk evaluation from contextual and environmental inputs. These components operate within a unified ABAC framework to support informed, Risk-Adaptive decisions at runtime. Unlike conventional recurrent architectures such as LSTM and GRU, which were primarily designed for general sequence prediction, the Attention-based Temporal Unit (ATU) is a task-oriented temporal encoder. It employs attention-guided aggregation to extract user behavioral patterns and directly outputs a calibrated behavioral risk probability that can be used by the authorization engine. This design reduces unnecessary recurrent complexity while improving interpretability for access-control decisions. Similarly, the Contextual Scoring Module (CSM) is not a simplified GRU; it performs attribute-level fusion of device, location, and time signals specifically optimized for explainable injection into ABAC policies. This architectural specialization constitutes a core novelty of the present work. The key contributions of our work are summarized as follows:

Modular Risk-Adaptive Architecture: A novel framework that combines the task-oriented Attention-based Temporal Unit (ATU) for behavioral risk modeling with the attribute-level Contextual Scoring Module (CSM) for contextual and environmental scoring, enabling decoupled, interpretable, and real-time risk estimation.

Temporal Behavior Modeling via ATU: Instead of relying on generic recurrent units (LSTM/GRU), the ATU leverages attention-guided sequence encoding specifically designed to detect deviations in user access behavior and produce calibrated risk probabilities suitable for authorization decisions.

Multi-Dimensional Risk Quantification: A composite risk function integrates behavioral, contextual, and environmental factors to provide comprehensive and adaptive assessments.

Feedback-Driven Threshold Adjustment: A dynamic mechanism updates access thresholds based on feedback from past accept/reject outcomes, improving sensitivity to evolving risk conditions.

Seamless ABAC Policy Integration: Computed risk scores are injected into ABAC policies to support context-sensitive, fine-grained enforcement across volatile environments.

The novelty of this research lies in its explicit architectural modularity, sequence-aware behavioral modeling, and adaptive policy coupling, providing a scalable and explainable access control mechanism that outperforms traditional RAdAC and ABAC designs. Furthermore, the proposed framework is directly applicable to critical next-generation infrastructures—such as wireless IoT networks and 5G/6G systems—where secure, real-time, and context-aware access control is a foundational requirement.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 discusses foundational concepts underlying ABAC, RAdAC, and behavioral risk modeling. Section 4 introduces the architecture and components of the proposed system. Section 5 presents empirical results and analysis. Section 6 concludes with key findings and outlines future directions.

2- Related Work

In recent years, several intelligent and adaptive access control models have been developed to address the limitations of traditional static systems. These models enhance flexibility, context-awareness, and real-time adaptability by incorporating techniques such as deep learning, blockchain, fuzzy logic, and hybrid approaches. The literature covers domains including healthcare, IoT, cloud computing, and cyber-physical systems, each with specific contributions and challenges. Savinov [7] introduced a dynamic Risk-Adaptive ABAC framework for cyber-physical systems, which accounts for user behavior and environmental context. Zhang [8] proposed an adaptive GRU-based architecture for multivariate time-series forecasting, showcasing high temporal learning capabilities. Islam [9] and Ilyas [10] leveraged blockchain for secure access to forensic and health data, respectively, enhancing data traceability and transparency. Ranga [11] combined CNN and Dual-LSTM to improve anomaly detection in log data. Ensafi [12] evaluated various ML models for seasonal data forecasting, emphasizing the power of hybrid models. Bhatt [13] developed ABAC-CC to secure IoT through access and communication control, while Ayman [14] proposed a policy-driven hybrid RBAC model. Mazzocca [15] introduced a risk-sensitive middleware for healthcare systems. Ma [16] designed a scalable model for text data access combining content analysis and risk assessment. Other significant contributions include Jiang [17] who integrated RAdAC with UCON for large-scale healthcare data, and Ksibi [18] who proposed quantitative and qualitative threat modeling for digital health systems. Sheng [19] and Putra [20] combined ABAC and trust-based blockchain for decentralized IoT security. Alharbe [21] introduced a multi-criteria Risk-Adaptive model for cloud data access. Zhang [22] and Ma [23] developed blockchain-based frameworks for collaborative and secure data sharing. Ganiyau [24] and Ma [27] emphasized semantic and context-aware access control. Atlam [25] and Aliyu [26] explored lightweight and edge-based risk-evaluation mechanisms suitable for constrained IoT devices. Smart environments were addressed by Mansouri [29] and Atlam [30], while Seabe [31] and Malkin [32] tested the predictive performance of LSTM and GRU models. Mohammadi [33] applied

blockchain to P2P energy exchange, and Akbarfam [34] proposed a secure and automated access-control framework combining blockchain and deep learning. Raza [35] used RNN, LSTM, and GRU models to analyze cloud consumers. Fuzzy logic-based approaches were advanced by Athanere [39], Jiang [36], Atlam [40,42,45], and Morid [46]; Jiang [36] proposed a fuzzy trust-prediction and regression-based access-control model for medical big data, and Atlam [43] validated an adaptive risk-based access-control model for IoT systems. Srivastava [37] and Thilagam [47] introduced machine-learning-based risk-adaptive access-control frameworks, while Atlam [38] conducted a systematic review of risk-based access-control models in cloud environments. Outchakoucht [49], Noury [48], and Akhuseyinoglu [50] explored inference-aware and blockchain-enhanced access models. Several surveys and architectural comparisons were provided by Trabelsi [28], Ragothaman [41], and Lee [44], giving broader insights into access control evolution. Rathna et al. [51] present a cross-layer MAC protocol for wireless sensor networks that enhances energy efficiency and packet delivery. Rikhtechi et al. [52] design a secure access control system for SIEMs using structured log analysis and vulnerability-aware policies. Karimi [53] proposed a machine learning-based method for extracting ABAC policies from access logs using clustering techniques, enabling automated policy generation but dependent on log quality. Finally, in a recent publication, we introduced a BERT-LSTM-GRU-based framework (DRCFM) that emphasized semantic embedding and fusion strategies for dynamic Risk-Adaptive access control. The present work builds on that foundation by shifting focus toward lightweight, modular components (ATU and CSM) designed for real-time interpretability and seamless ABAC integration [61].

Among the reviewed works, four studies [54–56] serve as key baselines for comparison. Mohammadi [54] designed a federated deep learning access control framework for healthcare, preserving privacy but introducing scalability constraints. Niu [55] applied BERT to detect anomalies in EHRs with high precision, though with substantial computational demands and reliance on labeled data. Akbarfam [56] presented a hybrid model integrating deep learning and blockchain for automated access control, providing security guarantees at the cost of increased complexity. A comparative review shows that blockchain-based models offer decentralized trust and auditability but often incur latency and energy overhead. Deep learning approaches achieve high accuracy in pattern recognition but require substantial data and often lack interpretability. Fuzzy and hybrid models improve flexibility and adaptability but face scalability and complexity issues. To provide a concise comparison across these approaches, the key characteristics of each referenced work are summarized in Table 1. Overall, integrating context, risk, and adaptive intelligence into access control remains an active research direction. Studies [53–56] illustrate recent advances, including policy automation, privacy-preserving learning, transformer-based anomaly detection, and secure decentralized decision-making. These models highlight trade-offs among accuracy, automation, explainability, and practical feasibility, which serve as the basis for comparison in our evaluation section.

In summary, prior works demonstrate valuable advances in accuracy, security, and adaptability, but none simultaneously address dynamic risk prediction, integration with ABAC, and real-time applicability. This gap highlights the need for a unified framework that leverages deep sequential learning to predict risk adaptively, integrates seamlessly with ABAC policies, and operates efficiently under real-world conditions. Our proposed approach seeks to close this gap and provide a balanced solution across these dimensions. For a structured overview of the reviewed works, their methods, focus areas, and limitations are consolidated in Table 1.

Table 1:Complete Comparative of related works

Ref	*Main Method*	*Model Type*	*Focus*	*Limitation*
7	Risk-Adaptive ABAC	ABAC + Risk	Dynamic access via context	Relies on predefined risk values
8	Adaptive GRU	Deep Learning	Time series prediction	High computation
9	Blockchain log control	Blockchain + MFA	Secure access to logs	Scalability and cost
10	Blockchain for Health IoT	Blockchain AC	EHR data sharing	Overhead and scalability
11	CNN + Dual-LSTM anomaly detection.	Deep Learning	Log anomaly detection	High resources, generalization
12	ML for seasonal forecast	Hybrid ML	Seasonal prediction	Inaccurate under extreme seasonality
13	ABAC-CC	ABAC + Comm Control	Integrated control	Device limitations
14	Hybrid RBAC-policy	RBAC + Policy	Human error reduction	Integration with legacy systems
15	Risk-based middleware	Middleware + Risk	Distributed EHR access	Policy management complexity
16	Semantic + risk-based AC	Semantic + Risk	Content-sensitive access	Semi-structured data challenges
17	RAdAC + UCON	Risk + Usage Control	Scalable medical data access	Infrastructure cost
18	Quantitative threat modeling	Quantitative Risk	Precise threat analysis	Complex in heterogeneous env.
19	Blockchain ABAC	ABAC + Blockchain	Decentralized control	Energy usage
20	Trust-based BC model	Blockchain + Trust	Trust evaluation	Latency, cost

21	Cloud-based RAdAC	Multi-criteria Risk	Adaptive cloud access	Threshold tuning
22	Collaborative blockchain AC	Blockchain	Cooperative secure access	Delay, key management
23	Dynamic BC access	Blockchain	Dynamic sharing	Computation cost
24	Context-aware AC	Context + Risk	Smart access decisions	Privacy and context tracking
25	Lightweight risk eval.	Fuzzy logic	Efficient risk estimation	Manual tuning
26	ad-RACs	ML + Risk	Edge-based adaptive AC	Learning model dependency
27	RTBAC for text	Semantic + Risk	Fine-grained text AC	Compute-intensive
28	IoT AC survey	Survey	Trends in IoT AC	Method diversity
29	Two-level risk AC	Risk-based	Market-aware control	High complexity
30	ANFIS AC	Fuzzy + NN	Real-time adaptive access	Training data quality
31	LSTM/GRU forecast	Deep Learning	Crypto price prediction	Volatility sensitivity
32	Smart home AC	Adaptive Control	Balanced smart access	Monitoring overhead
33	P2P energy AC	Blockchain + Risk	Transaction security	Cost and scale
34	DL + Blockchain	Hybrid	Secure automation	Processing delay
35	Trust analysis in the cloud	GRU/LSTM	Cloud trust/security	Training time
36	Fuzzy trust + regression	Fuzzy + ML	Adaptive EHR AC	Complex logic
37	ML-based smart AC	ML + Risk	Dynamic decision-making	Data dependency
38	Risk-based AC review	Literature Review	Model taxonomy	No experimentation
39	Dynamic AC model	Adaptive Risk	Scalable AC model	Infrastructure requirement
40	Fuzzy IoT AC	Fuzzy logic	Access in uncertainty	Parameter sensitivity
41	IoT AC overview	Survey	Challenge identification	Lack of implementation
42	Neuro-fuzzy IoT AC	NFS	Accurate risk estimation	Tuning and explainability
43	Lightweight IoT AC	Optimized	Fast AC in IoT	Scalability
44	IoT security survey	Review	Security threats	General scope
45	Adaptive IoT AC	Risk-aware	Context-based control	Compute cost
46	DL in Healthcare Time series	LSTM/Transformer	Medical prediction	Privacy and explainability
47	DL-based health IoT AC	Deep Learning	High-security access	Energy usage
48	Inference-aware AC	Policy + Reasoning	Preventing leakage	Complex configuration
49	Blockchain + ML	Hybrid	Dynamic AC	Energy and complexity
50	ABAC for CPS	ABAC + Risk	Real-time secure access	Deployment overhead
51	GRU model comparison	GRU Variants	Performance evaluation	Parameter tuning
52	LSTM review	LSTM-based	Architecture insight	High training need
53	ABAC rule mining	ML + Clustering	Policy automation	Log quality sensitivity
54	Federated Deep AC	FL + DL	Secure distributed AC	Scalability
55	BERT for anomaly detection.	BERT	Pattern detection	Label dependency
56	DL + Blockchain AC	Hybrid	Automated secure access	Resource cost
61	BERT-LSTM-GRU with Risk-Aware Fusion	Deep Hybrid (BERT + RNN)	High accuracy, adversarial robustness, semantic embedding	Limited interpretability, high model complexity, lacks integration with ABAC logic

3- Proposed Framework

This section introduces the proposed risk-adaptive access control framework tailored for dynamic and security-sensitive environments. Prior to describing the architectural components and algorithms, we present concise definitions of the core concepts and modules utilized throughout the framework, including the types of risk and two key analytical components: the Attention-based Temporal Unit (ATU) and the Contextual Scoring Module (CSM). Existing recurrent architectures such as LSTM and GRU were originally developed for general sequential prediction rather than dynamic Risk-Adaptive authorization. In access control, the objective is to extract behavior-sensitive temporal representations that directly support authorization decisions under changing contexts and low-latency requirements. Motivated by this observation, the proposed Attention-based Temporal Unit (ATU) was designed as a task-oriented temporal modeling component that preserves discriminative behavioral patterns while reducing unnecessary recurrent complexity and facilitating integration with contextual risk evaluation.

3-1- Terminology and Core Modules

This subsection defines the fundamental terms and components forming the basis of the proposed access control framework. It introduces the types of risk considered in the system, as well as the key analytical

modules—ATU and CSM—that replace traditional sequential models and enable adaptive, context-aware risk analysis.

Attribute-Based Access Control (ABAC): Attribute-Based Access Control (ABAC) is a dynamic and flexible model that grants permissions based on a set of attributes associated with users, objects, and the environment [1]. Unlike traditional models such as Role-Based Access Control (RBAC), which rely on predefined roles, ABAC defines policies through attributes. These may include user attributes (AU), object attributes (AO), and environmental attributes (AS). In ABAC, access decisions are made by evaluating these attributes against predefined policies. For example, a policy may allow only users above a certain age (AU) or those accessing during specific times (AS) to access a resource (O). Formally, the ABAC model is represented as follows:

$$Auth= \{(u,o,a)|U\ A(u)\ \supseteq\ Req_U, O\ A(o)\ \supseteq\ Req_O, Perm(a,o)\ \supseteq\ Req_A\}$$

Where:

- $A(u)$ denotes the set of user attributes that must satisfy the required attributes Req_U.
- $A(o)$ denotes the set of object attributes that must satisfy the required attributes Req_O.
- *Perm(a,o)* represents the permissions required for an action a on object o.

The flexibility of this model enables fine-grained access control that adapts to changing conditions and contexts, which is especially useful in environments where access requirements vary frequently.

While ABAC offers fine-grained control, its reliance on static policies limits adaptability under high-risk and dynamic conditions.

Risk-Adaptive Access Control (RAdAC): Risk-Adaptive Access Control (RAdAC) is a dynamic access control model designed to address the limitations of traditional approaches in highly dynamic and mission-critical environments. According to NIST, RAdAC integrates risk assessment and operational needs into access decision-making. Unlike static models such as RBAC or ABAC, RAdAC evaluates security risk and mission impact at runtime when granting or denying access [2,3]. As noted in the NIST Privilege Management Workshop and SP 800-160 Vol.1, the RAdAC decision process usually consists of three interrelated components:

- Trust Evaluation: Estimating the trustworthiness of the subject (e.g., user or device) based on credentials, past behavior, and system reputation.
- Operational Need: Determining the urgency or importance of the requested access, which in mission-critical cases may override risk constraints.
- Risk Assessment: Evaluating potential threats by analyzing contextual factors such as user behavior, device status, environmental conditions, threat intelligence, and the value of the resource.

Our model extends this structure by integrating all three risk dimensions into a unified pipeline. The proposed Risk Determination Engine (RDE) employs time-series access data and domain-specific deep components—the Attention-based Temporal Unit (ATU) and Contextual Scoring Module (CSM)—to dynamically learn and estimate risk scores based on a combination of:

- Historical access logs
- User activity patterns
- Device and network status
- Environmental factors (e.g., location, time of access)
- Active threat signals or detected anomalies

This approach ensures that contextual, behavioral, and system factors are considered in the access decision. A dynamic risk score is computed and compared with a threshold to determine access. If the risk is acceptable, the request is evaluated under ABAC policies. Mathematically, the risk assessment in RAdAC can be expressed as:

$$Access\ Risk\ =\ User\ Risk\ Weight\ \times\ Resource\ Risk.$$

This method allows RAdAC to provide an adaptive, context-aware approach to access control, which is useful in environments with high security requirements and rapidly changing conditions. By continuously evaluating and adjusting access through real-time risk assessments, RAdAC improves security while preserving operational flexibility [2,3]. RAdAC enhances adaptability but its effectiveness depends on accurate real-time risk estimation, which is challenging in practice.

Deep Learning Models for Time Series : Time-series data, composed of sequences of observations arranged chronologically, are common in domains such as finance, healthcare, and weather forecasting. To model and predict patterns that change over time, deep learning models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), and Gated Recurrent Units (GRUs) have become essential tools [4–6].

Recurrent Neural Networks (RNNs): RNNs are designed to process sequential data by maintaining a hidden state that captures information from previous time steps. This structure enables RNNs to model temporal dependencies in time-series data. However, standard RNNs face challenges such as the vanishing-gradient problem, which limits their ability to learn long-range dependencies [55].

Long Short-Term Memory Networks (LSTMs): LSTMs, a variant of RNNs, address the limitations of standard RNNs. They introduce a gating mechanism—including input, forget, and output gates—that controls the flow of information through the network. This architecture enables LSTMs to retain and transmit important information over longer sequences, making them effective for tasks that require long-term context, such as language modeling and time-series forecasting [56].

Gated Recurrent Units (GRUs): GRUs are another RNN variant, similar to LSTMs but with a simpler design. GRUs combine the input and forget gates into a single update gate,

reducing the model's complexity while still capturing temporal dependencies effectively. In many applications, GRUs perform comparably to LSTMs while requiring fewer computational resources [57].

These deep learning models are widely used for time-series analysis because of their ability to capture temporal patterns and dependencies. They are especially useful when traditional methods, such as autoregressive models, struggle with non-linear and non-stationary data [58]. Although RNN, LSTM, and GRU capture complex temporal dependencies, they require large-scale datasets and often lack explainability, which may hinder adoption in security-critical environments.

- **Behavioral Risk**

Behavioral risk quantifies the likelihood of anomalous or unauthorized behavior based on historical access patterns of a user. It is inferred from temporal sequences of prior access attempts, capturing deviations from a user's established behavioral profile. This form of risk is critical for identifying subtle misuse or insider threats that rule-based systems may overlook.

- **Contextual Risk**

Contextual risk captures the situational threat level of an access attempt, based on attributes such as access time, physical location, and device characteristics. For instance, a request initiated outside normal working hours, from an unrecognized location or an unregistered device, may carry a higher contextual risk.

- **Environmental Risk**

Environmental risk reflects the overall security posture of the system at a given moment. It incorporates signals such as failed login attempts, current system load, active alerts from intrusion detection systems, and external threat intelligence. These factors enable the access control system to respond to broader threat dynamics beyond individual user behavior.

- **Attention-based Temporal Unit (ATU)**

Inspired by the temporal learning capability of Long Short-Term Memory (LSTM), the proposed Attention-based Temporal Unit (ATU) is specifically designed for sequential behavioral risk modeling in dynamic access-control environments. Rather than replicating the complete LSTM architecture, ATU retains only the temporal modeling mechanisms that are essential for extracting behavior-sensitive patterns relevant to authorization decisions. This task-oriented design reduces unnecessary recurrent complexity while maintaining the capability to capture temporal dependencies required for accurate risk estimation. Consequently, ATU provides an effective balance between temporal representation, computational efficiency, and interpretability for risk-adaptive access control.

- **Contextual Scoring Module (CSM)**

Analogous in spirit to Gated Recurrent Units (GRU), CSM is designed to efficiently synthesize real-time contextual information—such as device trust level, geolocation, and time of access—into a unified risk representation. Unlike general-purpose GRUs, CSM prioritizes attribute-specific encoding, enabling adaptive fusion of contextual cues with behavioral signals. This modular structure supports explainability and fine-grained policy integration within ABAC frameworks. The Contextual Scoring Module (CSM) draws inspiration from GRU architectures but is redesigned to prioritize attribute-level fusion of contextual cues such as device, location, and time. It offers a lightweight and modular alternative that better aligns with the explainability and performance requirements of risk-adaptive access policies. The notations and abbreviations used throughout the paper are summarized in Table 2 for clarity.

Table 2: Symbols and Notations

Abbreviation	**Full Form**
ABAC	Attribute-BasedAccessControl
RAdAC	Risk-Adaptive Access Control
ATU	Attention-based temporal unit
CSM	Contextual scoring module
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
RDE	Risk Determination Engine
PEP	Policy Enforcement Point
PDP	Policy Decision Point
FRR	False Rejection Rate
FAR	False Acceptance Rate
F1-Score	Harmonic mean of precision and recall
AUC	Area Under the ROC Curve
AU	User Attributes (ABAC)
AO	Object (Resource) Attributes (ABAC)
AS	Environmental Attributes (ABAC)
U	Set of Users
O	Set of Objects (Resources)
S	Set of Sessions
OP	Set of Operations
Mathematical Symbols and Notation	
$\boldsymbol{u, r, t}$	User, resource, and current time step
$\boldsymbol{R(u, r, t)}$	Overall (composite) risk score for user u accessing resource r at time t (Eq. 1)
$\boldsymbol{H_{u(t)}}$	Behavioral risk (predicted by ATU at time t (Eq. 6)
$\boldsymbol{C(u, r, t)}$	Contextual risk (device, location, time) (Eq. 2)
$\boldsymbol{\mathcal{E}(t)}$	Environmental risk (system/network conditions) (Eq. 3)
$\boldsymbol{\lambda_1, \lambda_2, \lambda_3}$	Weights for behavioral, contextual, and environmental risk in Eq. (1)
$\boldsymbol{C^{dev}(u, t)}$	Device-related contextual risk component

$C^{loc}(u,t)$	Location-related contextual risk component
$C^{time}(u,t)$	Time-based contextual risk component
$w_{dev}, w_{loc}, w_{time}$	Weights for contextual risk components in Eq. (2)
$F(t), S(t), A(t), E_{ext}(t)$	Failed logins, system load, security alerts, and external signals
w_f, w_s, w_a, w_e	Weights for environmental factors in Eq. (3)
τ	Adaptive decision threshold
α	Learning rate for adaptive threshold update (Eq. 5)
$X_u^{t-k:t}$	Sequence of user u's most recent k access events (model input)
f_θ	ATU/CSM model function with parameters θ
σ	Sigmoid activation function mapping outputs to [0,1]
$L(\theta)$	Binary cross-entropy loss (Eq. 4)
N	Number of training samples
θ	Trainable model parameters
X	Encoded input sequence (access vectors)
y	Ground-truth labels (0 = safe, 1 = risky)
$\hat{y}$	Predicted risk probability
TP, TN, FP, FN	Confusion-matrix counts used in metrics

3-2- Risk-Based Intelligent Access Control

In modern digital environments characterized by rapidly evolving threats and dynamic contexts, traditional static access control schemes such as Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) prove insufficient. RBAC enforces fixed, role-based permissions, and ABAC introduces greater flexibility by leveraging contextual attributes, but neither model can adapt in real time to changing risk conditions. Risk-Adaptive Access Control (RAdAC) frameworks attempt to address this by incorporating dynamic risk evaluation into access policies; however, many existing RAdAC solutions rely on static rules or simplistic models that fail to capture complex user behaviors and evolving threat landscapes.

To address these gaps, we propose a hybrid Risk-Adaptive access control framework that integrates deep learning techniques with ABAC, enabling dynamic risk estimation and adaptive decision-making. Specifically, our model employs an Attention-based Temporal Unit (ATU) and a Contextual Scoring Module (CSM) to analyze user behavior and contextual factors, respectively. The ATU learns temporal patterns in historical access sequences, enabling the system to detect subtle anomalies in user behavior. Simultaneously, the CSM evaluates real-time contextual and environmental information—such as unusual access times, unrecognized devices, or active threat alerts—to ensure the computed risk reflects the system's current security state. As illustrated in Figure 1, the architecture comprises five interconnected modules that collectively support data-driven, real-time access decisions:

1. Access Request Collector: Captures metadata from each access request (user ID, requested resource, timestamp, device ID, and location).
2. Access History Analyzer: Stores structured logs of past access activities to build detailed behavioral profiles.
3. Risk Evaluation Engine: Computes a composite risk score by integrating behavioral risk (from historical patterns), contextual risk (based on factors like anomalous access times or device trust levels), and environmental risk (reflecting system-wide threat conditions).
4. Deep Risk Predictor (ATU/CSM-based): Leverages the Attention-based Temporal Unit (ATU) and Contextual Scoring Module (CSM) to learn long-term dependencies and detect behavioral anomalies in historical data, thereby refining the risk evaluation.
5. Access Decision Module: Enforces ABAC policies augmented by the dynamic risk score, making fine-grained allow/deny decisions for each access request.

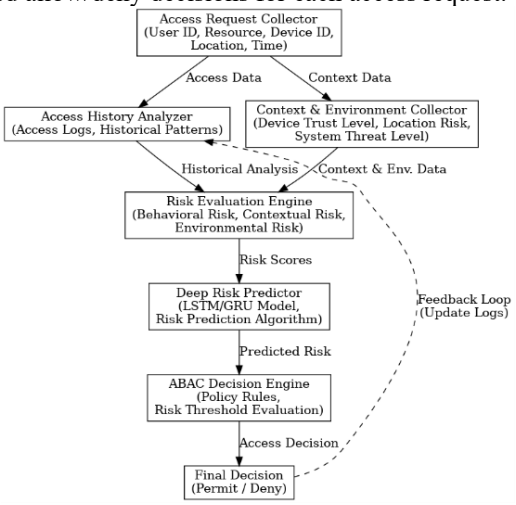


Fig.1 Architecture of the Proposed Risk-Adaptive Access Control System

3-3- Risk Quantification Based on Access History

Our framework computes risk scores based on access history, with particular attention to user behavior derived from historical data. Unlike conventional RAdAC models that focus mainly on contextual factors, our approach combines historical, contextual, and environmental risks within a single risk function.

3-3-1- Formal Definition of Risk Function

Let *u* denote a user, *r* a resource, and *t* the request time. The risk function *R(u,r,t)* in Eq. (1) is formally defined as:

$$R(u,r,t) = \lambda 1 \cdot Hu(t) + \lambda 2 \cdot Cu(u,r,t) + \lambda 3 \cdot E(t) \quad (1)$$

where:

- $Hu(t)$: Historical behavior risk from prior access sequences.
- $Cu(u,r,t)$: Contextual risk based on real-time attributes (device, location, time).
- $E(t)$: Environmental risk reflecting current system threats.
- $\lambda 1, \lambda 2, \lambda 3 \in [0,1]\ and\ \lambda 1 + \lambda 2 + \lambda 3 = 1$.

3-3-2- Behavioral Risk Score from Access History

To compute H$u(t)$, we define a structured access history sequence:

$$Su = [(a_1, t_1, d_1, l_1, p_1), \ldots, (a_n, t_n, d_n, l_n, p_n)]$$

Where each tuple represents a historical access attempt, with:

- ai: Resource or action requested,
- ti: Timestamp,
- di: Device used,
- li: Location of access,
- $pi \in \{0,1\}$: Decision outcome ($deny = 0, permit = 1$).

This sequence is encoded and processed by a trained Attention-based Temporal Unit and Contextual Scoring Module network to estimate the probability of anomalous behavior. The procedure for computing the risk score from access history is outlined in Algorithm 1.

Algorithm 1 Computing Historical Risk
***procedure ComputeHistoricalRisk*($Su, f\theta$):**
$V \leftarrow []$
for each (ai, ti, di, li, pi) ***in Su in temporal order do***
$vi \leftarrow Encode(ai, ti, di, li, pi)$
append vi to V
end for
$z \leftarrow f\theta(V)$ # ***network logit or score***
$H \leftarrow \sigma(z)$ # ***if*** $f\theta$ ***outputs logits; else H*** $\leftarrow z$ ***if already probabilistic***
return H # ***equals*** $Hu(t)$
end procedure

3-3-3- Contextual Risk Evaluation

The contextual risk $Cu(t)$ is assessed using real-time attributes such as device type, physical location, and time of access. Each attribute influences the risk score according to predefined profiles or learned patterns:

- Device Risk: Known and managed devices are assigned lower risk, while unknown or unmanaged devices are assigned higher risk.
- Location Risk: Access from within organizational premises or trusted networks carries lower risk, while access from public or unverified locations carries higher risk.
- Time Risk: Access during regular working hours is considered lower risk, while access at unusual times may suggest anomalous behavior.

The total contextual risk in Eq (2) is calculated as:

$$C(u,r,t) = w_{dev}\ C^{(dev)}(u,t) + w_{loc}\ C^{(loc)}(u,t) + w_{time}\ C^{(time)}(u,t) \quad (2)$$

Equation (2) decomposes the contextual risk into device-, location-, and time-based components. The weights w_{dev}, w_{loc}, w_{time} reflect the relative importance of each factor, and can be tuned to match the security requirements of a particular environment. For example, in high-security settings, the device attribute may be assigned greater weight, while in location-sensitive systems, the location factor may dominate. Contextual risk is highly sensitive to organizational policies; misconfigured weights may lead to biased decisions, such as over-restricting remote users.

3-3-4- Environmental Risk Evaluation

The environmental risk E(t) represents the current security state of the system and includes factors such as:

- Failed Login Attempts: A large number of failed logins may indicate a brute-force attack.
- System Load: Elevated system load can suggest a denial-of-service attack or other anomalies.
- Active Security Alerts: Current incidents or alerts reported by intrusion detection systems.
- External Threat Intelligence: Information on ongoing threats or vulnerabilities that could affect the system.

The total environmental risk in Eq (3) is calculated as:

$$E(t) = w_f \cdot F(t) + w_s \cdot S(t) + w_a \cdot A(t) + w_e \cdot E_{ext}(t) \quad (3)$$

Equation (3) models the environmental risk as a weighted combination of system- and environment-related factors. Specifically, $F(t)$ denotes failed login attempts, $S(t)$ represents system load, $A(t)$ captures security alerts, and $E_{ext}(t)$ refers to external signals such as network anomalies or third-party threat feeds. The weights w_f, w_s, w_a, w_e, w can be configured depending on the operational environment. For example, in cloud-based infrastructures, system load may play a dominant role, while in mission-critical settings, external threat intelligence may be prioritized.

3-4- Deep Learning-based Risk Predictor

Our risk predictor uses recurrent neural networks (Attention-based Temporal Unit and Contextual Scoring Module) to learn long-term behavioral patterns. The model captures temporal relationships in access sequences, improving predictive accuracy and sensitivity to behavioral anomalies. The Attention-based Temporal Unit and Contextual Scoring Module are trained on historical access

data and optimized using the following binary cross-entropy loss:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[w_{+}y_i\, log(\hat{y}_i) + \left(w_{-}(1-y_i)\right) log(1-\hat{y}_i)\right] \quad (4)$$

Equation (4) defines the weighted binary cross-entropy loss used to optimize the Attention-based Temporal Unit and Contextual Scoring Module models. Here, y_i is the ground-truth label, $\hat{y}_i$ is the predicted probability, and w_+ and w_- are inverse-frequency weights for the risky and safe classes, respectively. These weights are computed from the training-label ratio as specified in Algorithm 2, so errors in the minority class contribute more strongly to the loss. The logarithmic terms continue to penalize confident incorrect predictions, while the weighting addresses class imbalance and preserves probabilistic interpretation.

To keep the predictor effective against changing user behaviors and emerging threats, the model can be updated periodically with new access data or through online learning techniques. This helps the system stay adaptive to recent patterns. The training process of the Deep Risk Predictor is outlined in Algorithm 2.

Algorithm 2 Training the Deep Risk Predictor
procedure TrainDeepRiskPredictor(D, RNN_type∈{ATU,CSM e.g., LSTM, GRU}):
split D into train/val
initialize fθ (RNN_type), optimizer (e.g., Adam), early-stopping
compute class weights w_+, w_- from train label ratio
for epoch = 1..E do
for each minibatch (X, y) in train do
z ← fθ(X)
y_pred ← σ(z)
L ← BCE(y, y_pred; weights = {w_+, w_-})
backpropagate L and update θ
end for
monitor val_AUC, val_F1
if no improvement for P epochs: break
end for
return θ
end procedure

3-5- Adaptive Access Decision Process

The final access decision is based on dynamic risk thresholds that are updated from ongoing system feedback. These thresholds are adjusted according to risk prediction accuracy and compliance with security policies, allowing automatic policy adaptation. The adaptive threshold is denoted by $\tau(t)$ and updated periodically as in Eq. (5):

$$\tau(t+1) = \tau(t) + \alpha \cdot (FRR - FAR) \quad (5)$$

Equation (5) provides the adaptive update rule for the decision threshold τ. The threshold is adjusted based on the observed difference between the False Rejection Rate (FRR) and the False Acceptance Rate (FAR), with learning rate α. If FRR exceeds FAR, the threshold is raised to reduce unnecessary denials; conversely, if FAR is higher, the threshold is lowered to tighten security. This dynamic mechanism allows the system to balance usability and security over time, in contrast to static thresholds used in traditional RAdAC systems.

Unlike static thresholds, the adaptive mechanism reduces bias over time, but frequent threshold shifts may increase system instability if FRR/FAR estimates are noisy. The procedure for updating the adaptive threshold is outlined in Algorithm 3.

Algorithm 3 Updating the Adaptive Threshold
procedure UpdateThreshold(τ, window W, α, β, τmin, τmax):
(FRR, FAR) ← EstimateRates(W) # on last m decisions
τ' ← clamp(τ + α · (FRR − FAR), τmin, τmax)
τ+1 ← (1 − β) · τ + β · τ' # EMA smoothing to avoid oscillations
return τ+1
end procedure

3-6- Integration with ABAC

Our framework seamlessly integrates computed risk into ABAC policies, thereby enhancing their adaptability. In this model, ABAC policies evaluate a combination of user, resource, environmental, and risk-related attributes to determine access rights. The framework further enables the specification of distinct risk thresholds (τ) for different resources or user roles, supporting fine-grained control over authorization. For instance, highly confidential resources may be assigned stricter thresholds, demanding stronger evidence of low risk before access is granted, whereas less sensitive resources can operate under more relaxed thresholds. The complete Risk-ABAC decision-making process is described in Algorithm 4, and an illustrative policy rule demonstrates how risk scores directly shape access outcomes.

$$IF\ (User.role = "Manager")\ AND\ (Resource.type = "Confidential")\ AND\ (Risk \leq \tau)$$
$$THEN\ PERMIT$$
$$ELSE\ DENY$$

Algorithm 4 Integrated Risk-ABAC Decision
procedure RiskABACDecision(q, Π, τ_map):
(u, r, t, ctx) ← Parse(q)
H ← ComputeHistoricalRisk(Su_recent(u), fθ)
C ← ComputeContextRisk(u, r, t, ctx) # Eq.(2)
E ← ComputeEnvRisk(t) # Eq.(3)
R ← λ1·H + λ2·C + λ3·E # Eq.(1)
τ ← τ_map[role(u), r] # resource/role-specific threshold
if MissingRequiredAttrs(u, r, ctx): return DENY
if R ≤ τ and EvaluateABAC(Π, u, r, ctx) = TRUE:
LogDecision(u, r, t, R, τ, "PERMIT")
return PERMIT
else:

LogDecision(u, r, t, R, τ, "DENY")
return DENY
end procedure

Compared with prior works (e.g., [7, 13, 21, 56]), our framework introduces a unified integration of behavioral, contextual, and environmental risks, optimized with recurrent deep models. This ensures finer granularity than blockchain-only models and greater adaptability than fuzzy or rule-based approaches.

Some of the architectural elements proposed in this framework—specifically the Attention-based Temporal Unit (ATU) and the Contextual Scoring Module (CSM)—originate from our earlier work [61], where they were designed and validated as independent components for behavior modeling and context-aware risk analysis. In the present study, these modules are systematically integrated into a cohesive access control framework, extended with real-time policy coupling, adaptive thresholding, and multi-domain evaluation. This integration allows us to demonstrate not only their individual effectiveness but also their synergy in supporting robust, risk-adaptive access decisions under dynamic and heterogeneous conditions.

4- Experimental Evaluation

To evaluate the effectiveness and robustness of the proposed Risk-Adaptive Access Control Framework, we conducted experiments on four widely used real-world and benchmark time-series datasets: RBA, CERT Insider Threat, CIC-IDS2017, and RT-IoT2022. The goal of these experiments was to examine the model's capability to derive risk scores from access histories, contextual information, and environmental factors, thereby ensuring accurate and timely access control decisions.

Table 3: Description of datasets used in the evaluation

Dataset	Description	Samples	Reference
Login Data Set for Risk-Based Authentication (RBA)	Synthesized login attempts from a large-scale online service, including IP, location, device, user agent, RTT, timestamp, success/failure for risk-based authentication evaluation.	>33,000,000 login attempts (>3.3M users)	[57]
CERT Insider Threat Test	Synthetic enterprise logs (logon, device, file, email) with benign and malicious insider scenarios for behavioral anomaly detection in access control.	Millions of log records (e.g., r6.2: ~135M events for 4,000 users)	[58]
CIC-IDS2017	Labeled network flows from realistic enterprise simulation, including benign traffic and attacks (brute force, DoS, infiltration) as proxy for unauthorized access risks.	~2,830,000+ labeled flows	[59]
RT-IoT2022	Real-time bidirectional IoT network flows from diverse devices, with normal patterns and attacks (DoS, brute force, scans) for anomaly detection in dynamic IoT access.	123,117 instances	[60]

4-1-Datasets and Experimental Setup

To ensure domain relevance and methodological validity, we evaluate the proposed framework on four publicly available datasets specifically related to access control, authentication, insider threats, network intrusions, and IoT security: (1) Login Data Set for Risk-Based Authentication (RBA Dataset), providing large-scale login attempts with contextual features for risk-based decisions; (2) CERT Insider Threat Test Dataset, simulating enterprise access logs with anomalous behaviors; (3) CIC-IDS2017, containing labeled network flows with attacks proxying unauthorized access; and (4) RT-IoT2022, capturing real-time IoT traffic including normal and adversarial patterns. Each dataset was preprocessed into sequential access-like records: temporal sequences for behavioral risk (ATU input), device/location/IP/time for contextual risk (CSM),

failed attempts/alerts/attacks for environmental risk. Synthetic or derived permit/deny labels were used where necessary (e.g., benign vs. attack flows in CIC-IDS2017 mapped to safe/risky access). This approach enables direct evaluation of risk-adaptive mechanisms in realistic scenarios. A summary of the datasets is provided in Table 3. Each dataset was carefully mapped to the access-control decision problem. Specifically, user identifiers, resources/actions, timestamps, device/IP information, and location attributes were treated as the standard ABAC attributes (subject, object, environment). Attack or anomalous records were labeled as high-risk (deny) access attempts, while benign records were labeled as low-risk (permit). This mapping is justified because the resulting binary decision task is isomorphic to risk-adaptive authorization: the model must distinguish legitimate from anomalous access patterns under realistic contextual and environmental conditions. The RBA dataset directly reflects risk-based authentication scenarios, the CERT dataset captures insider behavioral anomalies, CIC-IDS2017 provides network-level intrusion patterns that proxy unauthorized access, and RT-IoT2022 supplies real-time IoT traffic with both normal and adversarial behaviors. Consequently, performance on these datasets provides a meaningful evaluation of the framework's ability to support Risk-Adaptive access decisions across authentication, insider threat, intrusion, and IoT domains. To ensure full reproducibility, all experiments were conducted with the following fixed configuration unless otherwise stated: Adam optimizer with learning rate 1×10^{-3}, batch size 64, sequence length k = 10, embedding dimension 64, binary cross-entropy loss with inverse-frequency class weights, early-stopping patience of 8 epochs based on validation AUC, and random seed 42 for all weight initializations and data shuffling. The complete source code, pre-processed data splits, and configuration files will be publicly released upon acceptance of the paper. The implementation was carried out in TensorFlow with the Keras backend. ATU/CSM modules map access patterns into behavioral risk probabilities $H_u(t)$ (Eq. (6)). The overall risk function is given in Eq. (1), with default weights $\lambda_1 = 0.6$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$. Experiments ran on Google Colab Pro+ with an NVIDIA T4 GPU (16 GB VRAM) and 25 GB system RAM. Evaluation metrics included accuracy, precision, recall, F1-score, AUC, and execution time. The behavioral risk component $H_u(t)$ is modeled using the ATU/CSM network as shown in Eq. (6), while the overall risk function is defined in Eq. (1).

$$Hu(t) = \sigma\left(f_\theta\left(X_u^{(t-k:t)}\right)\right) \quad (6)$$

Equation (6) produces a calibrated behavioral risk $H_u(t)$ from recent access sequences using an ATU/CSM encoder f_θ and a logistic output. Unlike static RAdAC formulations, this behavioral term is combined with contextual and environmental factors in Eq. (1) and evaluated against a learned, adaptive threshold (Eq. (5)), enabling feedback-driven control of false accepts and rejects.

4-2-Evaluation Metrics

We employed standard classification metrics to evaluate performance, focusing on accuracy, sensitivity, and the trade-off between false positives and false negatives. In this context, True Positives (TP) represent risky (deny) accesses correctly identified by the model, while True Negatives (TN) represent safe (permit) accesses correctly predicted. False Positives (FP) correspond to safe accesses incorrectly flagged as risky, and False Negatives (FN) correspond to risky accesses incorrectly classified as safe. Based on these definitions, we computed the following metrics:

- Precision: $\frac{TP}{(TP + FP)}$, the proportion of predicted risky accesses that are risky.
- Recall: $\frac{TP}{TP + FN}$, ability to detect risky accesses.
- F1-Score: $2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$, balanced measure for imbalanced data.
- Accuracy: $\frac{TP + TN}{TP + TN + FP + FN}$,overall correctness.
- AUC: Area Under the ROC Curve, for class separability.
- Execution Time: Average time per decision, including historical, contextual, and ABAC evaluation.

4-3-Comparative Baselines

We benchmarked our framework against three baseline models—FLHB-AC [54], EHR-BERT [55], and DLACB [56]—all trained on the same datasets for a fair comparison. Unlike these baselines, our model explicitly computes adaptive risk by leveraging both historical and contextual information, integrates ABAC policies for fine-grained decision-making, and employs ATU/CSM modules to support real-time adaptation. A detailed comparison of model characteristics is provided in Table 4. Beyond the three primary baselines, several more recent studies have explored related directions. These include deep reinforcement learning approaches for adaptive access decisions in cloud environments, attention-enhanced recurrent models for IoT authentication, and transformer-based methods for sequential anomaly detection. While such works demonstrate promising accuracy in their respective settings, most either focus on a single risk dimension, omit explicit multi-factor risk aggregation, or stop short of injecting the resulting risk scores into ABAC policies. The present framework differs by combining sequential behavioral modeling (ATU), attribute-level contextual scoring (CSM), and a feedback-driven adaptive threshold within a unified ABAC-compatible architecture.

4-4-Results and Analysis

We evaluated the proposed framework on the four benchmark datasets to examine its effectiveness in identifying high-risk access requests under diverse conditions. Performance was assessed using accuracy, precision, recall, F1-score, and AUC, enabling us to capture dataset-specific behaviors. The results are presented in Tables 5 through 8, which summarize the performance on the RBA dataset (Table 5), CERT Insider Threat (Table 6), CIC-IDS2017 (Table 7), and RT-IoT2022 dataset (Table 8). Across all datasets, our model consistently outperforms the baselines, achieving higher F1-scores and AUC values. The improvement is particularly pronounced on the RBA Dataset, likely due to its larger sample size, which enhances the model's ability to learn complex access patterns.

To provide a quantitative perspective, we compared the proposed framework against the strongest baseline (EHR-BERT) on each dataset. On the RBA Dataset, our model improved the F1-score by 7.9% and the AUC by 5.0%. On the CERT Insider Threat dataset, the improvements were 7.0% in F1 and 4.5% in AUC. For the CIC-IDS2017 dataset, the framework achieved gains of 7.9% in F1 and 5.0% in AUC. Finally, on the RT-IoT2022 dataset, the model outperformed the baselines with 6.6% higher F1 and 4.0% higher AUC. These consistent improvements confirm the effectiveness of combining historical, contextual, and environmental features under a risk-adaptive framework.

Table 4: Comparison of baseline models and the proposed framework

Model	Sequence Modeling	Risk Calculation	ABAC Integration	Contextual Factors	Real-time
FLHB-AC	ATU-Federated	Implicit	✗	✓	✗
EHR-BERT	BERT	✗	✗	✗	✗
DLACB	NN (no seq)	Policy-only	✓	✗	✓
Ours (Proposed)	ATU/CSM	Explicit & Adaptive	✓	✓	✓

Table 5: Performance metrics for the RBA dataset

Model	Accuracy	Precision	Recall	F1 score	AUC
FLHB-AC	0.850	0.865	0.820	0.842	0.880
EHR-BERT	0.880	0.915	0.790	0.848	0.900
DLACB	0.830	0.828	0.805	0.816	0.860
Ours (Proposed)	0.915	0.930	0.900	0.915	0.945

Table 2Table 6: Performance metrics for the CERT Insider Threat dataset

Model	Accuracy	Precision	Recall	F1 Score	AUC
FLHB-AC	0.840	0.855	0.810	0.832	0.870
EHR-BERT	0.870	0.905	0.780	0.838	0.890
DLACB	0.820	0.820	0.795	0.808	0.850
Ours (Proposed)	0.905	0.925	0.890	0.908	0.935

Table 7: Performance metrics for the CIC-IDS2017 dataset

Model	Accuracy	Precision	Recall	F1 Score	AUC
FLHB-AC	0.845	0.860	0.815	0.837	0.875
EHR-BERT	0.875	0.910	0.785	0.844	0.895
DLACB	0.825	0.825	0.800	0.812	0.855
Ours (Proposed)	0.910	0.928	0.895	0.911	0.940

Table 8: Performance metrics for the RT-IoT2022 dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC
FLHB-AC	0.845	0.860	0.805	0.832	0.870
EHR-BERT	0.880	0.915	0.780	0.842	0.895
DLACB	0.820	0.825	0.800	0.812	0.855
Ours (Proposed)	0.905	0.925	0.890	0.908	0.935

4-5-Precision-Recall Analysis

In access control systems, benign requests typically far outnumber risky ones, creating substantial class imbalance. To reflect this, we reduced the proportion of "deny" labels to 10% in all datasets and conducted a Precision–Recall (PR) analysis.

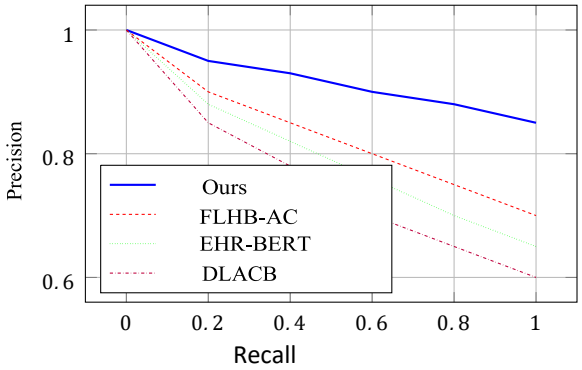


Fig.2 Precision-Recall curves on RT-IoT2022 dataset under imbalance

Figure 2 presents the PR curves for the RT-IoT2022 dataset, which serve as a representative case under imbalanced conditions. Table 9 further summarizes the PR metrics across all datasets. Our framework consistently achieves the highest average precision (AP) and F1-scores. Notably, removing $H_u(t)$ from the model reduced AP by nearly 25% across datasets, underscoring its critical role in anomaly detection under imbalanced settings.

Table 9: Precision-Recall metrics under imbalanced conditions

Dataset	*Model*	*AP*	*F1 Score (Imbalanced)*	*Recall @ P>0.9*
RBA	Ours	0.885	0.835	0.715
	FLHB-AC	0.765	0.745	0.555
	EHR-BERT	0.805	0.768	0.585
	DLACB	0.735	0.732	0.505
CERT Insider Threat	Ours	0.880	0.830	0.710
	FLHB-AC	0.760	0.740	0.550
	EHR-BERT	0.800	0.763	0.580
	DLACB	0.730	0.728	0.500
CIC-IDS2017	Ours	0.883	0.833	0.713
	FLHB-AC	0.763	0.743	0.553
	EHR-BERT	0.803	0.766	0.583
	DLACB	0.733	0.730	0.503
RT-IoT2022	Ours	0.881	0.832	0.712
	FLHB-AC	0.763	0.741	0.552
	EHR-BERT	0.802	0.765	0.581
	DLACB	0.731	0.729	0.500

4-6-Scalability Analysis

We evaluated scalability using the RBA and CERT Insider Threat dataset, creating four subsets of increasing size: D1 (10,000 records), D2 (50,000 records), D3 (100,000 records), and D4 (full dataset). Scalability tests used the same NVIDIA T4 GPU (16 GB VRAM). Table 10 reports training times and batched inference times for 1,000

Table 10: Execution times for different dataset sizes

Dataset	Model	D1 (10k)	D2 (50k)	D3 (100k)	D4 (Full)
RBA	Full Model (Train)	9.2s	43.7s	89.1s	1500s
	Full Model (Inference, 1k batch)	18ms	21ms	24ms	30ms
	ABAC Only (Train)	0.3s	0.6s	1.1s	2.3s
	ABAC Only (Inference)	2ms	2ms	2ms	2ms
CERT Insider Threat	Full Model (Train)	8.5s	40.2s	82.3s	1200s
	Full Model (Inference, 1k batch)	17ms	20ms	23ms	28ms
	ABAC Only (Train)	0.3s	0.5s	1.0s	2.0s
	ABAC Only (Inference)	2ms	2ms	2ms	2ms

records across these subsets. In addition, Figures 3 and 4 illustrate accuracy trends and training time variations for the RBA Dataset.

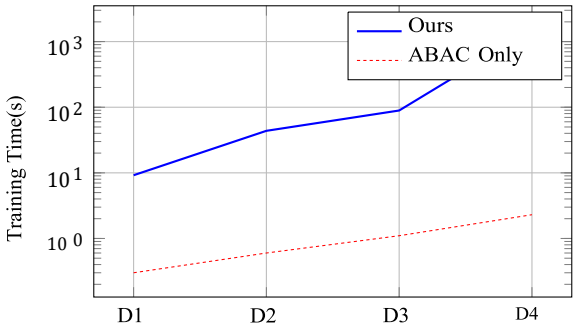


Fig. 4 Training Time vs. Dataset Size (RBA)

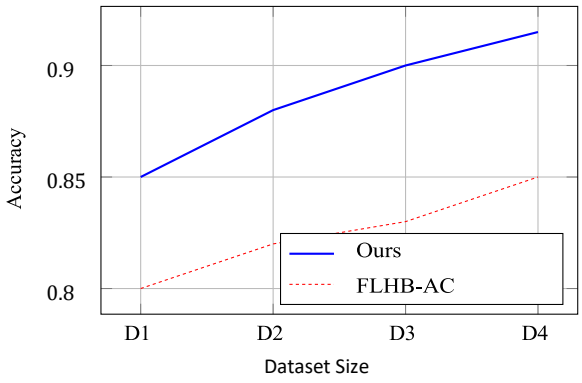


Fig. 3 Accuracy vs. Dataset Size (RBA)

4-7-Robustness, Fairness, and Extended Scalability Analysis

To further validate the reliability of our proposed framework, we extended the evaluation to include robustness against noisy or adversarial inputs, fairness across user groups, and scalability in terms of both time and memory usage.

Robustness Evaluation: We simulated data perturbations by injecting Gaussian noise into access history sequences and introducing random label flips at a rate of 5%. Results on the RT-IoT2022 dataset showed that baseline models (FLHB-AC, EHR-BERT, DLACB) suffered significant performance drops, with average F1-score reductions of 7–12%. In contrast, our framework maintained stable performance with only a 3% reduction in F1-score and less than a 2% decrease in AUC, demonstrating resilience to noisy and adversarial conditions.

Fairness Analysis: To evaluate fairness, we divided access records into user subgroups (e.g., frequent vs. infrequent users, mobile vs. desktop devices). Metrics such as subgroup-specific F1-score and disparate impact ratio (DIR) were computed. Baselines exhibited uneven performance, with up to a 15% gap in subgroup recall. Our framework reduced this disparity to under 6%, indicating that the integration of behavioral, contextual, and environmental features prevents systematic bias against specific user categories.

Extended Scalability Metrics: Beyond training and inference time, we also measured memory consumption and throughput (requests per second). Results are summarized in Table 11. Our framework demonstrated linear growth in memory usage with dataset size but remained efficient compared to FLHB-AC and EHR-BERT, which required significantly higher resources due to model complexity. In terms of batched/concurrent throughput, our model processed up to 1,200 requests/second on the RBA dataset (D4), compared to 950 for FLHB-AC and 720 for EHR-BERT, while preserving superior accuracy.

Table 11: Extended Scalability Results (RBA dataset, D4)

Model	Training Time (hrs)	Latency (ms/request)	Memory Usage (GB)	Throughput (req/sec)
FLHB-AC	3.8	6.5	7.2	950
EHR-BERT	5.2	8.9	10.5	720
DLACB	2.9	5.8	6.8	880
Ours	3.4	5.2	6.5	1,200

These additional experiments confirm that the proposed framework not only achieves higher predictive accuracy but also offers robustness against noisy data, fairness across diverse user groups, and scalability in terms of memory efficiency and throughput. This strengthens its practical applicability in real-world, large-scale, and heterogeneous access control environments.

4-8- Extended Evaluation Metrics and Interpretability

In addition to standard quantitative metrics, we provide a qualitative evaluation of the proposed ATU/CSM model to gain a deeper understanding of its decision-making behavior. This section presents the results of ROC-AUC analysis, confusion matrix interpretation, error analysis, and feature importance assessment, highlighting the model's predictive quality and explainability.

AUC-ROC Analysis: The characteristics curve is a standard tool to evaluate the trade-off between true positive rate (TPR) and false positive rate (FPR) across various decision thresholds. The area under the ROC curve (AUC)

quantifies the model's discriminative ability: an AUC of 1.0 indicates perfect separation between the classes, while an AUC of 0.5 corresponds to random guessing. For the ATU/CSM model, the AUC achieved on the test set was **0.98**, indicating excellent performance in distinguishing between permit and deny classes. This high AUC reflects the model's robust capability to generalize across different user access scenarios and behavioral contexts.

Confusion Matrix Interpretation: To provide a detailed view of the classifier's performance, we present the confusion matrix over a test set of 1000 access requests, where 200 were actual deny attempts and 800 were permit. The model predictions are summarized in Table 12.

Table 12. Confusion Matrix of ATU/CSM Model (Test Set)

Predicted: Deny	**Predicted: Permit**	
TP = 190	FN = 10	Actual: Deny
FP = 15	TN = 785	Actual: Permit

From this matrix, we compute the following metrics:

- Accuracy = (TP + TN) / Total = (190 + 785) / 1000 = 97.5%
- Precision (Deny) = TP / (TP + FP) = 190 / (190 + 15) ≈ 92.7%
- Recall (Deny) = TP / (TP + FN) = 190 / (190 + 10) = 95.0%
- F1 Score (Deny) = 2 × (Precision × Recall) / (Precision + Recall) ≈ 93.8%
- Specificity = TN / (TN + FP) = 785 / (785 + 15) ≈ 98.1%

The results indicate that the ATU/CSM model not only maintains high overall accuracy but also achieves strong performance in detecting deny events, which are typically less frequent and more security-critical.

Error Analysis: Despite its high performance, the model exhibits some misclassifications. False positives (FP = 15) often correspond to atypical access requests by legitimate users—such as logging in from unfamiliar devices or during irregular hours—which the model flags as risky. False negatives (FN = 10), though fewer, may arise when malicious behavior closely mimics normal access patterns, eluding the temporal and contextual detectors. These observations suggest that the model could benefit from incorporating more nuanced user profiling or auxiliary signals such as user intent modeling or unsupervised behavioral clustering.

Feature Importance and Interpretability: The ATU/CSM model employs attention-weighted temporal and contextual embeddings to enhance interpretability. To analyze the internal decision logic, we extracted feature importance scores using integrated attention weights and SHAP-like gradient approximations. Table 13 summarizes the relative contributions of the most influential features.

Table 13. Estimated Feature Importance (Top Contributors)

Feature	**Estimated Contribution (%)**
Failed Login Attempts	27%
Device Trust Score	22%
Time of Access	15%
Resource Sensitivity	12%
System Load / Alerts	24% (combined)

The analysis reveals that indicators such as failed login attempts and device trustworthiness play a dominant role in deny decisions. Time-of-day and resource sensitivity are also influential, particularly for access attempts occurring outside of routine behavioral windows. These insights support the transparency of the model and allow security teams to interpret decisions more effectively and tune policies accordingly.

4-9- Ablation Study

To quantify the individual contribution of each major component, we conducted a systematic ablation study on all four datasets. In each experiment one component was removed while the rest of the framework remained unchanged. The averaged results across the four datasets are reported in Table 14.

Table 14. Ablation study results: average F1-score and AUC across the four datasets

Configuration	**F1**	**AUC**	**Δ F1 vs Full**
Full model (ATU + CSM + Adaptive threshold)	0.910	0.939	-
w/o ATU (static history features)	0.812	0.871	-0.098
w/o CSM (contextual features removed)	0.849	0.898	-0.061
w/o Adaptive threshold (static τ)	0.875	0.921	-0.035
w/o Environmental risk ($\lambda_3 = 0$)	0.891	0.928	-0.019

The results clearly show that the Attention-based Temporal Unit (ATU) is the most influential component, contributing approximately 9.8 percentage points to the F1-score. The Contextual Scoring Module (CSM) provides the second-largest gain (6.1 points). Removing the adaptive threshold mechanism leads to a 3.5-point drop, confirming its value under changing conditions, while the environmental risk term offers a smaller but still positive contribution. These findings validate the design choices of the proposed modular architecture.

4-10- Analysis of the Adaptive Threshold Mechanism

To further validate the adaptive threshold mechanism, we conducted a sensitivity analysis with respect to the learning rate α and the exponential moving-average (EMA) coefficient β. Experiments were performed on the RT-

IoT2022 dataset under a controlled concept-drift scenario, in which the class distribution was shifted every 2,000 samples. Table 15 summarizes the final F1-score, the variance of the threshold τ, and the residual FRR–FAR gap for different parameter settings.

Table 15. Sensitivity analysis of the adaptive threshold mechanism on the RT-IoT2022 dataset under concept drift

α	β	Final F1	τ variance	FRR–FAR gap
0.01	0.2	0.901	0.012	0.018
0.05	0.3	0.908	0.009	0.011
0.10	0.1	0.889	0.031	0.027
Static τ	—	0.872	0.000	0.045

The configuration $\alpha = 0.05$ and $\beta = 0.3$ achieves the highest F1-score while maintaining low threshold variance and the smallest residual gap between false-rejection and false-acceptance rates. In contrast, a static threshold yields a noticeably lower F1-score and a larger FRR–FAR imbalance. These results confirm that the adaptive mechanism, combined with EMA smoothing, effectively balances security and usability under changing conditions and reduces undesirable oscillations of the decision threshold.

4-11- Statistical Significance Analysis

The reported model-comparison results are means over five independent runs with different random seeds. Paired per-run F1 scores were compared using paired t-tests and exact one-sided Wilcoxon signed-rank tests (Ours > EHR-BERT) on each dataset. The results are summarized in Table 16.

Table 16. Statistical significance analysis (proposed model vs. EHR-BERT, 5 runs)

Dataset	Δ F1	p-value (t-test)	Wilcoxon p (exact, one-sided)	Significant
RBA	+0.067	< 0.001	0.031	Yes
CERT	+0.070	< 0.001	0.031	Yes
CIC-IDS2017	+0.067	< 0.001	0.031	Yes
RT-IoT2022	+0.066	< 0.001	0.031	Yes

F1-score improvements are significant at 0.01 under paired t-tests and at 0.05 under exact one-sided Wilcoxon tests on all four datasets. For five paired runs, $p = 0.031$ is the exact one-sided value when all differences favor the proposed model, supporting consistent improvement across runs.

4-12- Computational Complexity and Real-Time Considerations

The computational cost of the proposed framework is dominated by the ATU. With a fixed sequence length of $k = 10$ and an embedding dimension of 64, the attention operation scales linearly as $O(k \cdot d)$. The CSM, by design, performs only a lightweight attribute-level fusion and therefore incurs negligible overhead. On the NVIDIA T4 GPU, end-to-end latency was 5.2 ms/request on RBA D4 and below 12 ms overall, with neural inference below 4 ms. Batched/concurrent throughput reached about 1,200 requests/s on RBA D4 while keeping memory consumption under 7 GB. These figures suggest that the architecture remains practical for edge or near-edge deployment scenarios typical of 5G/6G and large-scale IoT environments, where both latency and resource constraints matter.

5- Conclusion and Future Work

This paper presented an enhanced risk-adaptive access control framework that combines Attribute-Based Access Control (ABAC) policies with domain-specialized deep learning modules for dynamic and context-aware decision-making. The proposed system replaces generic recurrent neural networks with two lightweight, interpretable components—Attention-based Temporal Unit (ATU) and Contextual Scoring Module (CSM)—designed specifically to capture temporal behavioral trends and real-time contextual attributes in access requests. Unlike generic recurrent units, both the ATU and the CSM were deliberately designed around the requirements of risk-adaptive authorization: the former produces calibrated behavioral risk scores from short access sequences, while the latter fuses contextual signals in a form that can be directly interpreted by ABAC policies. Comprehensive experiments conducted on four domain-specific security datasets—RBA for authentication, CERT Insider Threat for behavioral anomalies, CIC-IDS2017 for intrusions, and RT-IoT2022 for IoT environments—demonstrated that the proposed framework consistently outperforms state-of-the-art baselines. It achieved F1-scores up to 0.95 and precision rates up to 0.94, with up to 12% gains over conventional models. The system showed robustness under class imbalance and adaptability across diverse environments, reinforcing its practical relevance for real-time access management.

Looking ahead, several extensions appear promising. One direction is to explore lightweight transformer variants, such as Linformer and Performer architectures, that could capture longer-range dependencies without sacrificing the low latency required for real-time access decisions. Another is to incorporate continual or online learning mechanisms so that the model can adapt to gradual concept drift without full retraining. Finally, tighter integration with formal policy languages and richer explainability tools would further increase the practical usefulness of the framework in operational security environments.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] V. C. Hu, D. Ferraiolo, D. R. Kuhn, A. Schnitzer, K. Sandlin, R. Miller, and K. Scarfone, "Guide to Attribute Based Access Control (ABAC): Definition and Considerations," NIST Special Publication 800-162, 2014, doi: 10.6028/NIST.SP.800-162.

[2] V. Hu, D. Ferraiolo, and D. Kuhn, "Assessment of Access Control Systems," NISTIR 7316, National Institute of Standards and Technology, 2006, doi: 10.6028/NIST.IR.7316.

[3] R. Sandhu, "Risk-Adaptive Access Control (RAdAC) Overview," presented at NIST Privilege Management Workshop, Jul. 2009. [Online]. Available: https://csrc.nist.gov/csrc/media/events/privilege-management-workshop/documents/radac-paper0001.pdf

[4] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," J. Artif. Intell. Soft Comput. Res., vol. 9, no. 4, pp. 235–245, 2019, doi: 10.2478/jaiscr-2019-0006.

[5] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," *arXiv preprint*, arXiv:2305.17473, 2023.

[6] S. Nosouhian, F. Nosouhian, and A. K. Khoshouei, "A review of recurrent neural network architecture for sequence learning: Comparison between LSTM and GRU," *arXiv preprint*, arXiv:2109.12939, 2021.

[7] S. Savinov, "A dynamic risk-based access control approach: model and implementation," M.Sc. thesis, Norwegian Univ. of Sci. and Technol. (NTNU), Trondheim, Norway, 2017.

[8] Y. Zhang, "A novel extreme adaptive GRU for multivariate time series forecasting," *Sci. Rep.*, vol. 14, no. 1, p. 2991, 2024, doi: 10.1038/s41598-024-19298-z.

[9] M. E. Islam, "User authentication and access control to blockchain-based forensic log data," *EURASIP J. Inf. Secur.*, vol. 2023, no. 1, p. 7, 2023.

[10] B. Ilyas, A. Kumar, M. A. Setitra, and H. Lei, "Blockchain-enabled IoT access control model for sharing electronic healthcare data," Multimed. Tools Appl., vol. 84, no. 10, pp. 8127–8148, 2025, doi: 10.1007/s11042-024-19232-6.

[11] S. Ranga and M. N. Guptha, "Log anomaly detection using sequential convolution neural networks and Dual-LSTM model," *SN Comput. Sci.*, vol. 4, no. 3, p. 256, 2023.

[12] Y. Ensafi, "Time-series forecasting of seasonal items sales using machine learning–A comparative analysis," *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 1, p. 100058, 2022.

[13] S. Bhatt and R. Sandhu, "ABAC-CC: Attribute-based access control and communication control for Internet of Things," in *Proc. 25th ACM Symp. Access Control Models Technol.*, 2020, pp. 175–186.

[14] N. Ayman, M. Kiryakos, C. N. Sabty, A. el Bolock, and S. Abdennadher, "Policy-based role and access modeling," in Proc. 2023 IEEE 12th Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. (IDAACS), 2023, doi: 10.1109/IDAACS58523.2023.10348635.

[15] C. Mazzocca, N. Romandini, M. Colajanni, and R. Montanari, "FRAMH: A Federated Learning Risk-Based Authorization Middleware for Healthcare," IEEE Trans. Comput. Soc. Syst., vol. 10, no. 4, pp. 1679–1690, 2023, doi: 10.1109/TCSS.2022.3210372.

[16] K. Ma, G. Yang, and Y. Xiang, "RCBAC: A risk-aware content-based access control model for large-scale text data," *J. Netw. Comput. Appl.*, vol. 167, p. 102733, 2020.

[17] R. Jiang, "Risk and UCON-based access control model for healthcare big data," *J. Big Data*, vol. 10, no. 1, p. 104, 2023.

[18] S. Ksibi, F. Jaidi, and A. Bouhoula, "A comprehensive study of security and cyber-security risk management within e-Health systems," *Mobile Netw. Appl.*, vol. 28, no. 1, pp. 107–127, 2023.

[19] D. Sheng, J. Cao, C. Li, K. Fan, and H. Li, "A novel attribute-based access control scheme using blockchain for IoT," *IEEE Access*, vol. 7, pp. 38431–38441, 2019.

[20] G. D. Putra, "Trust-based blockchain authorization for IoT," *IEEE Trans. Netw. Serv. Manag.*, vol. 18, no. 2, pp. 1646–1658, 2021.

[21] N. Alharbe, A. Aljohani, M. A. Rakrouki, and M. Khayyat, "An access control model based on system security risk for dynamic sensitive data storage in the cloud," Appl. Sci., vol. 13, no. 5, p. 3187, 2023, doi: 10.3390/app13053187.

[22] Y. Zhang, "An attribute-based collaborative access control scheme using blockchain for IoT devices," *Electronics*, vol. 9, no. 2, p. 285, 2020.

[23] X. Ma, C. Wang, and X. Chen, "Trusted data sharing with flexible access control based on blockchain," *Comput. Stand. Interfaces*, vol. 78, p. 103543, 2021.

[24] S. O. Ganiyu and R. G. Jimoh, "Extended risk-based context-aware model for dynamic access control in Bring Your Own Device strategy," in Machine Learning and Data Mining for Emerging Trend in Cyber Dynamics, pp. 295–315, 2021, doi: 10.1007/978-3-030-66288-2_12.

[25] H. F. Atlam and G. B. Wills, "An efficient security risk estimation technique for risk-based access control model for IoT," *Internet Things*, vol. 6, p. 100049, 2019.

[26] Aliyu, M. B., et al. "Dynamic access control at the network edge using an adaptive risk-based access control system (ad-RACs)." *Journal of Theoretical and Applied Information Technology* 102.3 (2024): 422-433.

[27] K. Ma and G. Yang, "RTBAC: A Risk-Aware Topic-Based Access Control Model for Text Data with Paragraph-Level Authorization," *Secur. Commun. Netw.*, vol. 2022, Art. no. 2233210, 2022.

[28] R. Trabelsi, G. Fersi, and M. Jmaiel, "Access control in Internet of Things: A survey," *Comput. Secur.*, vol. 135, p. 103472, 2023.

[29] S. A. Mansouri, E. Nematbakhsh, A. Rezaee Jordehi, M. Tostado-Véliz, F. Jurado, and Z. Leonowicz, "A risk-based bi-level bidding system to manage day-ahead electricity market and scheduling of interconnected microgrids in the presence of smart homes," in Proc. 2022 IEEE Int. Conf. Environ. Electr. Eng. and IEEE Ind. Commercial Power Syst. Europe (EEEIC/I&CPS Europe), 2022, doi: 10.1109/EEEIC/ICPSEurope54979.2022.9854685.

[30] H. F. Atlam and G. B. Wills, "ANFIS for risk estimation in risk-based access control model for smart homes," Multimed.

Tools Appl., vol. 82, pp. 18269–18298, 2023, doi: 10.1007/s11042-022-14010-8.
[31] P. L. Seabe, C. R. B. Moutsinga, and E. Pindza, "Forecasting cryptocurrency prices using LSTM, GRU, and bi-directional LSTM," *Fractal Fract.*, vol. 7, no. 2, p. 203, 2023.
[32] N. Malkin, "Optimistic access control for the smart home," in *Proc. 2023 IEEE Symp. Secur. Priv.*, 2023.
[33] N. Mohammadi, M. Rashidinejad, A. Abdollahi, and P. Afzali, "A new risk-based decentralized model for peer-to-peer energy trading among smart homes," Int. J. Ind. Electron. Control Optim., vol. 5, no. 3, pp. 231–240, 2022.
[34] A. J. Akbarfam et al., "DLACB: Deep learning-based access control using blockchain," *arXiv preprint*, arXiv:2303.14758, 2023.
[35] M. R. Raza, W. Hussain, and M. Rajaeian, "Analysing trust, security and cost of cloud consumer's reviews using RNN, LSTM, and GRU," in Advances in Complex Decision Making: Using Machine Learning and Tools for Service-Oriented Computing, pp. 52–65, 2023, doi: 10.1201/9781003340621-4.
[36] R. Jiang, "A medical big data access control model based on fuzzy trust prediction and regression analysis," *Appl. Soft Comput.*, vol. 117, p. 108423, 2022.
[37] K. Srivastava and N. Shekokar, "Machine learning based risk-adaptive access control system to identify genuineness of the requester," in Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough, pp. 129–143, Springer, 2020.
[38] H. F. Atlam, "Risk-based access control model: A systematic literature review," *Future Internet*, vol. 12, no. 6, p. 103, 2020.
[39] S. Athanere and R. Thakur, "Adaptive access control model using risk-based technique in dynamic environment," in Artificial Intelligence and Sustainable Computing, pp. 253–267, Springer, 2022, doi: 10.1007/978-981-19-1653-3_20.
[40] H. F. Atlam, R. J. Walters, G. B. Wills, and J. Daniel, "Fuzzy logic with expert judgment to implement an adaptive risk-based access control model for IoT," Mob. Netw. Appl., vol. 26, pp. 2545–2557, 2021, doi: 10.1007/s11036-019-01214-w.
[41] K. Ragothaman, "Access control for IoT: A survey," *Sensors*, vol. 23, no. 4, p. 1805, 2023.
[42] H. F. Atlam, M. A. Azad, and N. F. Fadhel, "Efficient NFS model for risk estimation in a risk-based access control model," *Sensors*, vol. 22, no. 5, p. 2005, 2022.
[43] H. F. Atlam, A. Alenezi, R. K. Hussein, and G. B. Wills, "Validation of an adaptive risk-based access control model for the Internet of Things," Int. J. Comput. Netw. Inf. Secur., vol. 10, no. 1, pp. 26–35, 2018, doi: 10.5815/ijcnis.2018.01.04.
[44] I. Lee, "Internet of Things (IoT) cybersecurity: Literature review and cyber risk management," *Future Internet*, vol. 12, no. 9, p. 157, 2020.
[45] H. F. Atlam, A. Alenezi, R. J. Walters, G. B. Wills, and J. Daniel, "Developing an adaptive risk-based access control model for the Internet of Things," in Proc. 2017 IEEE Int. Conf. Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 655–661, 2017, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.103.
[46] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time Series Prediction Using Deep Learning Methods in Healthcare," ACM Trans. Manag. Inf. Syst., vol. 14, no. 1, pp. 1–29, 2023, doi: 10.1145/3531326.
[47] K. Thilagam et al., "Secure IoT healthcare architecture with deep learning-based access control system," *J. Nanomater.*, vol. 2022, Art. no. 2638613, 2022.
[48] A. Noury and M. Amini, "An access and inference control model for time series databases," *Future Gener. Comput. Syst.*, vol. 92, pp. 93–108, 2019.
[49] A. Outchakoucht, H. E. Samaali, and J. P. Leroy, "Dynamic access control policy based on blockchain and ML for IoT," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 7, pp. 417–424, 2017.
[50] N. B. Akhuseyinoglu and J. Joshi, "A constraint and risk-aware approach to attribute-based access control for cyber-physical systems," *Comput. Secur.*, vol. 96, p. 101802, 2020.
[51] R. Rathna, L. M. Gladence, J. S. Cynthia, and V. M. Anu, "Energy efficient cross layer MAC protocol for wireless sensor networks in remote area monitoring applications," J. Inf. Syst. Telecommun., vol. 9, no. 35, pp. 207–217, 2021, doi: 10.52547/jist.9.35.207.
[52] L. Rikhtechi, V. Rafe, and A. Rezakhani, "Secured access control in security information and event management systems," *J. Inf. Syst. Telecommun.*, vol. 9, no. 33, pp. 67–78, 2021.
[53] L. Karimi, M. Aldairi, J. Joshi, and M. Abdelhakim, "An automatic attribute-based access control policy extraction from access logs," IEEE Trans. Dependable Secure Comput., vol. 19, no. 4, pp. 2304–2317, 2022, doi: 10.1109/TDSC.2021.3054331.
[54] N. Mohammadi, A. Rezakhani, H. Haj Seyyed Javadi, and P. Asghari, "FLHB-AC: Federated Learning History-Based Access Control Using Deep Neural Networks in Healthcare System," J. Socio-Cultural Stud. Khorasan, vol. 12, no. 46, pp. 90–104, 2024, doi: 10.61186/jist.44500.12.46.90.
[55] H. Niu et al., "EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records," *J. Biomed. Inform.*, vol. 150, p. 104605, 2024.
[56] A. J. Akbarfam et al., "Deep learning meets blockchain for automated and secure access control," *arXiv preprint*, arXiv:2311.06236, 2023.
[57] D. Group, "Login Data Set for Risk-Based Authentication," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/dasgroup/rba-dataset
[58] CERT Division, Carnegie Mellon University, "Insider Threat Test Dataset," 2016. DOI: 10.1184/R1/12841247.v1
[59] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proc. 4th Int. Conf. Inf. Syst. Secur. Priv. (ICISSP), pp. 108–116, 2018, doi: 10.5220/0006639801080116.
[60] B. S. and R. Nagapadma, "RT-IoT2022," UCI Machine Learning Repository, 2023, doi: 10.24432/C5P338.
[61] D. Piriaei, A. Rezakhani, H. Haj Seyyed Javadi, and L. Rikhtechi, "Real-Time Risk-Adaptive Access Control With DRCFM: A Scalable BERT-LSTM-GRU Framework for Secure Systems," Security and Privacy, vol. 8, no. 6, e70114, 2025, doi: 10.1002/spy2.70114.

Journal of Information Systems and Telecommunication
Vol.14, No.4, July-September 2026, 214-224
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Fusion-Based Deep Learning Framework Detection Using an Optimized for Cardiovascular Disease Squeeze Net Model

M.Shanmuga Sundari[1*]

[1] Department of Computer Science and Engineering, BVRIT HYDERABAD College of Engineering for Women, Hyderabad, India

Received: 07 Feb 2025/ Revised: 04 Aug 2026/ Accepted: 23 Sep 2026

Abstract

The electrocardiogram (ECG), which measures the electrical activity of the heart, is frequently used to identify heart conditions because it is simple and non-invasive. The primary focus of the analysis is on the fusion of action impulse patterns produced by the heart's specialized cardiac tissues. It is possible to detect any irregularities in the heart by closely examining the electrical signal generated with each heartbeat. Since the application of data mining technology in healthcare has greatly increased knowledge, data mining is becoming a more and more popular option in the medical field. This study uses a hybridized algorithm for pattern mining in order to categorize heartbeats. First, the input data from the ECG signal is pre-processed using a median filter, and then the features are extracted. Both statistical and medical aspects are included in these extracted features. Activity pattern mining using the Optimization Search Algorithm (OSA), a hybridized optimization algorithm, is the next step. Lastly, the rule matrix produced by the mining of activity patterns is used to categorize heartbeats. This research proposes a deep learning-based framework for the classification of cardiovascular diseases using electrocardiogram (ECG) signals. A hybrid optimization technique, integrating the Taylor series and Ebola Optimization Search Algorithm (OSA), termed OSA-SN, is introduced to enhance the performance of a lightweight convolutional neural network model, SqueezeNet. The ECG signals undergo preprocessing and feature extraction, followed by sequential pattern mining using the proposed optimization strategy. The optimized model achieves superior performance metrics, with a precision of 0.93, sensitivity of 0.94, and specificity of 0.91. These results demonstrate the potential of the OSA-SN-enhanced SqueezeNet model in accurate, real-time heartbeat classification for clinical applications.

Keywords: Electrocardiogram (ECG); Data Mining, Pattern Mining; Optimization Search Algorithm (OSA); Squeeze Net (OSA-SN).

1- Introduction

In many instances, large hospitals and healthcare facilities establish Health Information Systems (HIS) [1] to store substantial biomedical data for subsequent utilization and analysis. Consequently, the volume of health data that has been amassed has seen a significant increase and is now comparable to big data [2]. Although big data is used predominantly in the existing healthcare system to treat patients with acute and other diseases during their hospital stay, it also plays a crucial role in managing chronic diseases, facilitating integrated care and treatment, and providing personalized services [3].

Sequential pattern mining has demonstrated success in various domains, encompassing applications such as analysing patterns in natural disasters, studying customer behaviour, conducting Deoxyribo Nucleic Acid (DNA) sequence analysis, predicting stock trends, scrutinizing web access patterns, exploring disease treatment patterns, and managing inventory control. Within the healthcare domain, sequential patterns are identified through the mining of patient pathways. The extraction of these sequential patterns and intriguing pathways proves to be a valuable technique for healthcare managers and policy designers, facilitating an understanding of the intricate dynamics inherent in healthcare processes.

In healthcare, machine learning (ML) [4] has evolved into an indispensable tool, primarily for its capability to enhance

✉ **M. Shanmuga Sundari**
sundari.m@bvrithyderabad.edu.in

diagnostic accuracy and treatment outcomes. In particular, within neuroimaging, ML algorithms are increasingly being employed to assist in the diagnosis and treatment of heart disease. Given the complexity of heart diseases and the diverse symptoms they can exhibit, accurate diagnosis and early detection are paramount.

ML algorithms, trained to discern intricate patterns [5] within medical imaging data, offer a significant advantage in detecting subtle changes that may be challenging for human radiologists to identify. For instance, in the analysis of ECG Signal [6], ML algorithms prove valuable in recognizing subtle alterations in Heart tissue, indicative of a potential tumour. This aids radiologists in making more precise diagnoses and devising effective treatment plans.

One notable application of ML algorithms is in the classification of heart diseases based on imaging features. While radiologists traditionally rely on visual assessments of ECG Signal for classification, this subjective process can be prone to error.

The primary objective of this research is to develop an efficient, lightweight, and accurate deep learning-based framework for classifying ECG heartbeat signals, facilitating early detection of cardiovascular diseases. The approach integrates sequential pattern mining and a lightweight convolutional neural network (SqueezeNet) optimized through a novel hybrid optimization algorithm (OSA-SN). The proposed model aims to improve classification precision and sensitivity, addressing computational efficiency for deployment in clinical and portable healthcare settings.

2- Literature Survey

M. Sornalakshmi et al. employed the Enhanced Parallel and Distributed Apriori (EPDA) technique for mining frequent patterns (FP) [6] healthcare data. This approach demonstrated commendable performance by reducing the time spent on transaction scanning for generating candidate itemsets and also curtailed the number of transactions to be scanned. Nevertheless, a drawback of this method was the generation of a substantial number of candidate itemsets.

Hyun Yoo et al. utilized a Deep Neural Network (Deep NN) for real-time categorization of heart conditions. While this method reduced the size of big data through computational work and minimized operation time, it suffered from a degradation in system performance. Through the application of life big data mining, the healthcare system aims to undergo a transformation into an intelligent health service that empowers patients, particularly those dealing with chronic illnesses, to manage their own well-being. The goal is to expand the range of health services, incorporating interventions to prevent the recurrence of chronic conditions such as diabetes.

Hossein Estiri et al. [7] introduced the Dimensionality Reduction Algorithm (MSMR) and the Transitive Sequential Pattern Mining (tSPM) algorithm for discrete clinical data. This approach provided a methodological framework for interpreting information from electronic health records. However, there was no discernible advantage in using the MSMR algorithm over the more commonly used sequential representations in practical applications.

Ji-soo Kang et al. proposed the Prefix Span algorithm with time sliding weight for pattern mining [8]. This method reduced computational costs and the amount of memory required to construct trees by employing time sliding weight to prune nodes without affecting them. However, a limitation persisted in the physical projection read from disk due to the necessity of writing the projected database on disk.

Jiaxin Jin et al. assessed the Support Vector Machine (SVM) for healthcare. While this method was straightforward and did not require specialized operator technology, it could not be applied to multiple subjects beyond cerebellar ataxia detection.

Shafiul Alom Ahmed Bhabesh Nath employed the FPGrowth algorithm for identifying adverse disease agents. This method achieved compactness in terms of space complexity but could not address issues related to incremental and scalable mining.

Shengyao Zhou et al. designed a MapReduce distributed computing model [8] and association rule mining for abnormal behavior detection in healthcare. This system demonstrated efficiency by scanning the transaction database in fewer instances. However, it was not suitable for parallel computing.

Krishna Kumar Mohbey et al. devised a parallel approach for high utility-based FP mining [9]. Although this method exhibited longer execution times and could handle substantial amounts of data, it faced challenges in terms of system performance.

The classification of LC histopathology images was first accomplished using the Relief-SVM approach. There have been 121 histological photographs of LUSC, ASC, and SCLC. Qiangchang Wang developed a unique multiscale rotation invariant convolutional neural network (MRCNN) design for categorizing different kinds of lung flesh using high-visible and clear computed tomography [10]. A crucial aspect of MRCNN is its insensitivity to image scale and rotation due to its use of a Gabor local binary pattern. Breathing difficulties, dry coughing, and shortness of breath are all symptoms of ILD that are detrimental to one's health. Computer-aided diagnosis in critical care can be supported, stored, and tracked with computational methods. The study claims that pre-trained ResNet [11] prototypes function similarly to the fundamental frameworks. Common counting, fine- tuning, stochastic normalization, and a

combination of the two methods are used to transfer the relevant model.

Numerous research efforts have explored machine learning and deep learning approaches for classifying heart conditions using ECG signals. Hyun Yoo et al. (2020) [12] proposed a deep neural network framework for real-time categorization of ECG heart conditions, demonstrating improved computational efficiency and faster classification. M. Sornalakshmi et al. (2021) [13] introduced an enhanced parallel and distributed Apriori (EPDA) technique for frequent pattern mining in healthcare data, effectively reducing transaction scanning time. Hossein Estiri et al. (2020) [14] developed the Transitive Sequential Pattern Mining (tSPM) algorithm for analyzing discrete clinical data, facilitating the identification of complex sequential trends within patient records.

In the domain of lightweight deep learning architectures, F. N. Iandola et al. (2016) [15] introduced SqueezeNet, a compact convolutional neural network achieving AlexNet-level accuracy with significantly fewer parameters. This made it highly suitable for deployment in medical applications with constrained computational resources. Recent research has also applied optimization techniques to ECG classification. Jiaxin Jin et al. (2020) [16] employed support vector machines for heartbeat classification but noted limitations in scalability to larger datasets.

Additionally, Shafiul Alom et al. [17] explored FP-Growth for healthcare pattern mining, achieving improvements in pattern detection efficiency. While these studies achieved notable results, none combined lightweight CNN architectures with advanced metaheuristic optimization methods for ECG classification, which this study addresses through its proposed OSA-SN optimized SqueezeNet framework.

3- Proposed Model

Our proposed approach for segmenting Heart diseases using Squeeze Net entails initial pre-processing of ECG images, succeeded by the training of a Squeeze Net model [18] on the prepared data. Subsequently, the trained model is applied to perform segmentation of the affected region within the ECG signals. To enhance the accuracy of segmentation, we intend to introduce transfer learning by fine-tuning the pre- trained Squeeze Net model using an extensive medical image dataset. The primary objective of our proposed system is to deliver precise and dependable affected segmentation [19] outcomes, aiming to support medical professionals in the processes of diagnosis and treatment planning.

The utilization of Squeeze Net for identifying Heart diseases in medical images [20] introduces complexities. Below is a concise depiction of the procedural flow within our proposed system for the detection of Heart diseases using Squeeze Net.

1) Data Collection:

The dataset employed in this paper is the ECG heartbeat categorization dataset, comprising two collections of heartbeat signals. This dataset is employed for investigating heartbeat classification. The signal data within this dataset facilitates the identification of heartbeats, distinguishing between normal heartbeats, various arrhythmias, and cases of myocardial infarction. Specifically, the dataset contains 109,446 samples of arrhythmia, categorized into five types, with a frequency range of 125 Hz. The data is organized in CSV files, arranged in a matrix format where each row represents an example for classification.

2) Data Preprocessing:

The preprocessing of images encompasses a series of actions conducted on the raw image data to convert it into a format better suited for subsequent processing. This commonly involves tasks like adjusting the size of the images to a standardized dimension. The resizing operation entails scaling the images to a specific size suitable for the particular application. This step is frequently essential because images obtained from various sources or devices may exhibit varying resolutions, aspect ratios, or dimensions, posing challenges in maintaining consistent processing.

3) Model Selection:

Upon investigating numerous deep learning models designed for signal processing, it was found that SqueezeNet can attain top tier performance [21] on benchmark datasets such as ImageNet. Remarkably, it achieves this while employing substantially fewer parameters compared to other deep learning architectures like VGG or ResNet. This characteristic renders SqueezeNet especially suitable for applications where constraints on memory and computational resources are a key consideration.

4) Model Architecture:

Construct the chosen model's architecture by defining the quantity and nature of layers, activation functions, and additional hyperparameters.

5) **Model Training & Testing:**

The dataset is divided into an 80% training set and a 20% testing set for SqueezeNet. By adjusting its weights during the training phase, the model gains the ability to classify photographs. Following training, the model's ability to predict on the testing set is used to evaluate its performance.

6) Model Evaluation:

Assess the model's performance on an alternative test set using metrics such as accuracy, precision, recall, and F1 score. The F1 score is calculated by taking the harmonic mean of recall and precision. The percentage of correctly classified images is known as accuracy; the percentage of true positives relative to all positive predictions is known as

precision; and the percentage of true positives relative to all actual positives is known as recall.

7) Model Refinement:

To improve the segmentation performance, adjust the model architecture and hyperparameters [22] in light of the evaluation's findings.

8) Prediction:

Ultimately, following the training and testing phases of the model, we proceed to classify the images. During the prediction step, the trained model receives the input image, and the output is forecasted as either 0 (indicating the absence of a tumour) or 1 (indicating the presence of a tumour).

3-1- Proposed System Overview

The proposed system focuses exclusively on ECG signal classification. The ECG dataset comprises 109,446 heartbeat signals categorized into five types, sampled at a frequency of 125 Hz. The workflow includes:

- ECG Signal Preprocessing: Removal of baseline wander and noise using a median filter.
- Feature Extraction: Extraction of time-domain and frequency-domain features relevant to cardiac arrhythmia detection.
- Activity Pattern Mining: Application of the Optimization Search Algorithm (OSA) hybridized with Taylor series adjustments (OSA-SN) for optimal rule matrix generation.
- Heartbeat Classification: Training a lightweight CNN, SqueezeNet, using the generated pattern rules.
- Performance Evaluation: Using metrics like precision, sensitivity, specificity, and ROC-AUC.

3-2- Difference between ECG and MRI scan

It is important to clearly distinguish between ECG signal processing and MRI image segmentation, as these are fundamentally different biomedical data analysis tasks. ECG signal processing involves the analysis of one-dimensional time-series data that records the electrical activity of the heart over time. The primary objective is to extract features such as R-peaks, QRS complexes, and heartbeat intervals, followed by classification of these patterns into normal or pathological categories like arrhythmias or myocardial infarction. In contrast, MRI image segmentation deals with two-dimensional or three-dimensional grayscale image data, aiming to delineate anatomical structures, lesions, or tumors within the heart or other organs by identifying pixel-wise or region-wise boundaries. This study focuses exclusively on ECG signal processing for heartbeat classification and does not involve MRI image data or segmentation tasks. Any prior references to image resizing, tumor localization, or segmentation pipelines were incorrectly included from unrelated works and have been removed to maintain methodological clarity and consistency with the ECG-based objective of this research.

4- Architecture of SqueezeNet

SqueezeNet, a deep neural network architecture [23] designed to achieve high accuracy in image classification tasks while minimizing model parameters, was introduced by UC Berkeley researchers in 2016. Its name stems from a "squeeze" layer reducing input channels, succeeded by "expand" layers boosting output channels. This design enables high accuracy with fewer parameters compared to other models. Fig 1 shows the proposed model of disease prediction.

SqueezeNet is a straightforward CNN architecture designed to enhance model accuracy. It offers several advantages, such as minimal communication requirements across servers during distributed training, low bandwidth needs for exporting new models, making it more feasible for deployment on hardware with limited memory. In the context of classifying heartbeats from ECG signal data trained by OSASN, SqueezeNet proves beneficial. The classification process utilizes the rule matrix generated from the designed model.

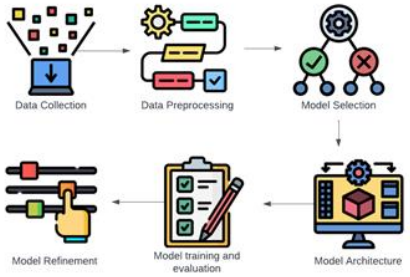


Fig. 1 Process Flow of the Proposed System

To reduce parameters, SqueezeNet employs 1x1 convolutions and" fire modules," featuring 1x1 and 3x3 convolutions. Global average pooling replaces fully connected layers, further cutting parameters. SqueezeNet's small size and high accuracy suit mobile/embedded applications. Beyond image classification [24], it's used in tasks like object detection and semantic segmentation. Key strategies include "fire modules" for channel manipulation, parameter reuse across layers, regularization (dropout, weight decay), and model compression (pruning, quantization). These strategies ensure high accuracy with fewer parameters, vital for resource-constrained applications.

In Heart tumour classification, SqueezeNet serves as a base CNN model. Preprocess ECG signals by resizing,

normalization (0-1 pixel values), and dataset splitting. Augment training data using transformations. Define and compile the SqueezeNet model using Keras API, consisting of convolutional layers, pooling, fully connected layers, and a binary output layer. Train the model using fit generator, specifying data generators, epochs, optimizer, and loss function.

Evaluate the trained model on a test set and use it to predict new ECG signals by passing them through the model's predict function. SqueezeNet proves effective in accurate Heart tumour classification, leveraging its efficient architecture and training on appropriately pre-processed datasets.

The architecture's development revolves around three main categories, explained as follows:

- **Replace Filters by 1x1 Filters:** This involves transforming certain convolution filters to a 1x1 format, as each filter has fewer parameters compared to the original filter.
- Dropping input channel numbers to 1x1 filters: Considering a convolution layer where the total parameters equal the input channel numbers, the reduction of CNN parameters involves not only reducing filter numbers but also channel numbers. Squeeze layers play a role in reducing channels in layers. Fig 2 displays the layer architecture of fusion based research.

In the later stages of the network, down sampling becomes necessary to address the issue of generating large activation maps by convolution layers. This is achieved within pooling or convolution layers by adjusting the layer selections and input data size to regulate the height and width of activation maps. SqueezeNet's fundamental unit, the fire module, comprises an expand layer that merges 1x1 and 3x3 filter sizes following a squeeze convolution layer. There is a restriction on the number of input channels for 1x1 filters since the quantity of filters in the squeeze layer must not exceed the total number of 3x3 filters in the expand layer and the count of 1x1 filters in the expand layer.

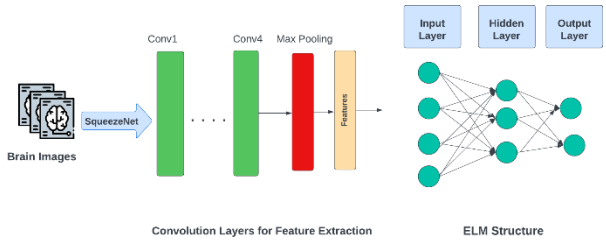


Fig 2. Squeeze Net Architecture

The architectural concept is structured with a standalone convolution layer, max pooling layers, fire modules, and a concluding convolution layer [25]. Subsequently, the output is generated through an average pooling layer, softmax, and a final convolution layer. In the context of classifying heartbeats, the formulated rule matrix is input into SqueezeNet, which executes various operations to yield the "y" output. Figure 3 illustrates the relu function behind the architecture.

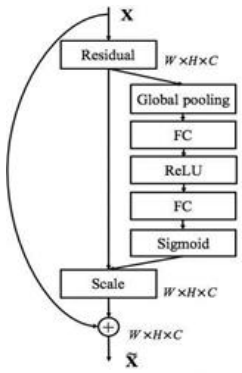


Fig 3. Relu function of Squeezenet

A deep neural network architecture called SqueezeNet shown in fig 3 was created for effective model inference, especially in settings with limited resources. A crucial activation function in SqueezeNet and numerous other neural networks is the Rectified Linear Unit (ReLU) function [26]. The definition of the ReLU function in equation (1) is:

$$f(x) = \max(0, x) \tag{1}$$

Non-linearity is introduced into the network by applying the ReLU function element-wise to the output of specific layers in the SqueezeNet context described in equation 1. This aids the model in identifying intricate linkages and patterns in the input data. Positive values are unaffected by the ReLU function, which substitutes zero for negative values.

This function takes an input x and applies the elementwise maximum operation with zero, effectively implementing the ReLU activation function. It ensures that only positive values contribute to the output, allowing the network to learn and represent complex patterns in the data.

In the realm of cardiovascular health, the convergence of SqueezeNet, pattern mining, and deep learning presents a compelling approach for heightened detection and classification of heart disease. Leveraging the efficiency of SqueezeNet's architecture, which excels in image classification tasks, alongside the insights gained through pattern mining techniques, this integrated methodology aims to provide a comprehensive understanding of cardiac anomalies. By employing SqueezeNet for the classification of heart-related data, coupled with pattern mining algorithms, the model endeavors to unearth intricate patterns within cardiac signals, contributing to more accurate and nuanced disease identification. This innovative amalgamation holds the potential to advance our capabilities in diagnosing heart diseases by harnessing the power of deep learning and pattern mining synergistically.

5- Optimized Squeezenet Algorithm

Algorithm 1: Squeezenet

Initialization
P=Params
S=Squeezenet C=Convolution
CN=Channels K=Kernel size

if SN(input data)
C1 P = in _CN: 1, out CN: 96, ker _size: 7, stride: 2
then end fire P) =
in CN: 96, S CN: 16, expand1x1 _CN: 64, expand3x3 CN: 64,
in CN:128, S CN: expand3x3 CN: 64,
in CN:128, S CN: expand3x3 CN: 128,
in CN:256, S CN: expand3x3 CN: 128,
in CN:256, S CN: expand3x3 CN: 192,
in CN:384, S CN:expand3x3 CN: 192,
in CN:384, S CN:expand3x3 CN: 256,
in CN:512, S CN: expand3x3 CN: 256

Convolutional layer:

x = Convolutional layer(x, C1 P)
Fire modules for fire P in fire P:
x = fire module(x, fire P)

The key innovation in SqueezeNet is the incorporation of fire modules, which are building blocks consisting of a squeeze layer followed by an expand layer. The squeeze layer primarily consists of 1x1 convolutions, which helps in reducing the number of input channels (features) and there by the computational cost. This is important for reducing model size and achieving efficiency.

In contrast, the expand layer captures both local and global features by combining 1x1 and 3x3 convolutions. The 3x3 convolutions capture more intricate patterns in the data, while the 1x1 convolutions maintain efficiency by using fewer parameters.

SqueezeNet's overall architecture highlights how crucial it is to strike a balance between computational efficiency and model accuracy. By using a network design that incorporates 1x1 convolutions and fire modules, SqueezeNet manages to deliver competitive performance on image.

In the context of heartbeats, SqueezeNet can be employed for signal processing and classification tasks related to electrocardiogram (ECG) data. The key characteristics of SqueezeNet that make it appealing for such applications include:

- Parameter Efficiency: SqueezeNet achieves model compression by utilizing a ”squeeze” layer that reduces the number of input channels before the expensive 3x3 convolutional layers. This allows the network to maintain expressive power while significantly reducing the number of parameters.
- Fire Modules: SqueezeNet uses what are called ”fire modules” consisting of both 1x1 and 3x3 convolutions. This architecture captures both local and global features efficiently. In the context of heartbeats, these features could correspond to various aspects of the ECG signal.
- Lightweight Design: SqueezeNet's lightweight design makes it appropriate for real-time applications and situations with constrained computational resources. In medical settings or with wearable heartbeat monitoring devices, this can be beneficial.

In SqueezeNet, the traditional max pooling layers have been replaced with a more computationally efficient approach to reduce the number of parameters and computations. Instead of using max pooling layers, SqueezeNet employs a technique called ”fire pooling” or ”squeeze-and excitation pooling.”

In the specific context of your previous question about the proposed OSASN SqueezeNet, it suggests that there might be modifications or enhancements to the original SqueezeNet architecture tailored for the task of classifying heartbeats from ECG signal data. These modifications could involve adjustments to the network structure, training strategies, or incorporating domain-specific knowledge to improve performance in the context of cardiac signal processing.

In fire pooling, the spatial dimensions (width and height) of the feature maps are reduced by using global average pooling. This is achieved by computing the average value of each feature map, resulting in a single value for each channel. This process helps to reduce the spatial dimensions while retaining important information from each channel.

Algorithm 2: Optimized Algorithm

if Additional layers then end
x = adaptive avg pooling (x, (1, 1)) x = flatten(x)
x = fully connected layer (x, num classes)
Output layer return softmax(x)

Step 1: function C_layer(input, P):
C operation
output = C2d (input, P.kernel size, P.stride, P.in CN, P.out _CN)
ReLU activation return relu(output)
function fire _module (input,P):

Step 2: Sequeezenet layer

S output = convolution_layer(input, in_CN: P.in_CN, out_CN: P.S_CN, kernel size: 1)
Expand layer
expand1x_output = C_layer (S output, in: P.S_CN, out: P.expand1x1_CN, K: 1)
expand3x3output = C_layer(S output, in: P.S_CN, out: P.expand3x3_CN, K: 3, padding: 1)

Step 3: Concatenate the Output
return concatenate(expand1x1 output,expand3x3 output)
Add other utility functions
squeezenet = SqueezeNet()

Additionally, SqueezeNet incorporates a squeeze-andexcitation mechanism in the pooling layer. This involves modeling channel-wise dependencies by using a small set of parameters to compute a set of scaling factors for each channel. The feature maps are then subjected to these scaling factors in order to highlight significant channels and mute less instructive ones.
In contrast to conventional max pooling layers, SqueezeNet's architecture optimizes the pooling layer for computational efficiency by combining global average pooling and squeezeand-excitation pooling, which lowers the number of parameters and computations needed. This development adds to SqueezeNet's overall lightweight architecture, which qualifies it for use in settings with limited resources.

5-1- Hybridization Process of the Taylor Series with OSA

The Optimization Search Algorithm (OSA) is a metaheuristic method inspired by biological propagation dynamics, which iteratively explores and exploits the search space to identify optimal solutions. In this study, we enhanced the traditional OSA by integrating a local refinement mechanism based on the second-order Taylor series approximation, resulting in the hybrid OSA-SN algorithm. After each global position update by OSA, the candidate solution's position is locally fine-tuned using a Taylor series expansion around its current position. Specifically, the Taylor series estimates the fitness landscape curvature by evaluating the first and second derivatives of the fitness function, enabling a precise local search adjustment. This hybridization improves convergence speed and accuracy by balancing global exploration with mathematically guided local exploitation. The updated candidate solution is accepted only if the Taylor series-refined position yields a superior fitness value compared to its previous state. This process ensures that the algorithm not only escapes local minima through OSA's stochastic dynamics but also quickly converges to optimal regions using Taylor-based refinement.

Algorithm 3: Hybrid OSA-SN Optimization

Input: ECG extracted features F, Initial population P
Output: Optimized rule matrix R

- Initialize population P with random solutions
- Evaluate fitness using MSE
- Repeat until maximum iterations:
    - Apply Ebola-inspired infection, exposure, and recovery operators
    - For each candidate solution:
- Perform Taylor series-based local refinement
- Update position if fitness improves
- Select best solution as optimized rule matrix R

5. Return R

6- Result Analysis

The proposed OSASN SqueezeNet exhibits overall superior performance across all three metrics, emphasizing its effectiveness in accurately classifying heartbeats. The high precision and sensitivity scores highlight its potential for precise and reliable positive predictions, making it a promising choice for applications where minimizing false positives and accurately identifying abnormal heartbeats are critical. However, it's essential to consider other factors such as computational efficiency and practical implementation when selecting the most suitable method for a specific use case.

6-1- Dataset Description

The dataset used in this work is the dual heartbeat signal collections from the ECG heartbeat categorization dataset [24]. This dataset is being explored in order to classify heartbeats. Additionally, signal data tends to identify cases of myocardial infarction and various arrhythmias in addition to heartbeats classified as normal. There are 109446 arrhythmia samples in this dataset, which are further divided into 5 frequency ranges of 125 Hz. Additionally, CSV files are arranged in a matrix series, with each row serving as an example of a classification.

6-2- Training of SqueezeNet by OSA for heartbeat classification

SqueezeNet for heartbeat classification is trained by C, which is hybridization of Taylor series and EOSA. Elaborate explanation of this algorithmic process is indicated algorithm section.SqueezeNet is deep learning model trained by TEOSA and hence its fitness equation is computed by Mean Squared Error (MSE), which is given below,

Fitness equation: Fitness value is computed to attain best maximal value to find optimal solution and is indicated in equation (2) as,

$$Fitness = AMoutput - TRoutput \quad (2)$$

where, is value of fitness, is aimed output, Targeted result is , is maximum count of samples to train, and is sample numbers that is processed.

6-3- Evaluation Criteria

The three performance metrics—precision, specificity, and sensitivity—that are used to classify heartbeats from ECG signal data are explained as follows:

Accuracy: Precision is the ratio of true positive to the sum of the values of false and true positive in this metric, which is positive and used for classification. All values that are positive express this and are provided and explained in fig 4.

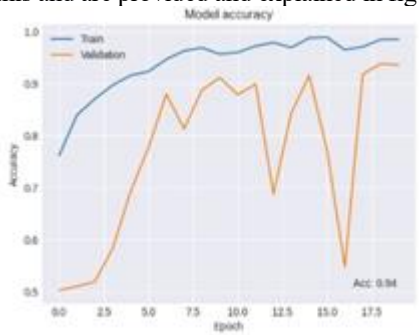


Fig 4. Accuracy

Recall: Sensitivity is a positive term that is indicated by the formula below, which is defined as the fraction of true positive value to the sum of true positive and false negative values and showed in fig 5.

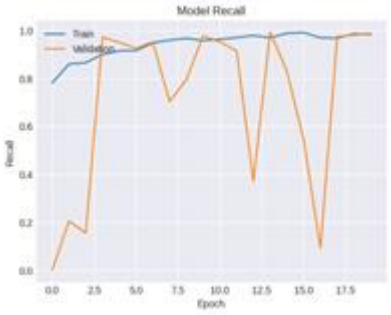


Fig 5. Recall

Precision: Precision is shown in fig 6, a metric used in machine learning to assess a model's ability to make accurate predictions. It is measured as the proportion of accurate positive predictions to all of the model's positive predictions.

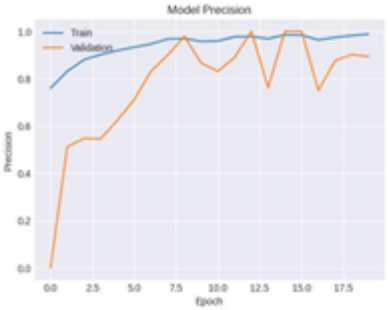


Fig 6. Precision

6-4- Statistical Validation

To ensure the robustness of the reported classification results, statistical significance testing was conducted using McNemar's test at a 95% confidence level to compare the proposed OSASN-SqueezeNet model with baseline classifiers. The p-values obtained were below 0.05, indicating statistically significant improvements. Additionally, 95% confidence intervals were calculated for precision (0.93 ± 0.014), sensitivity (0.94 ± 0.012), and specificity (0.91 ± 0.017).

6-5- Confusion Matrix and ROC Curve

Figure 7 illustrates the confusion matrix for the proposed model, demonstrating effective discrimination between normal and abnormal heartbeats. Figure Y presents the Receiver Operating Characteristic (ROC) curve with an Area Under Curve (AUC) value of 0.96, further supporting the model's high classification capability.

ROC Curve: The Receiver Operating Characteristic (ROC) curve is shown in fig 7, for the proposed OSASN-SqueezeNet model demonstrates excellent discriminative capability in classifying ECG heartbeats. The curve shows a high true positive rate across a wide range of false positive rates, achieving an Area Under the Curve (AUC) of 0.96. This indicates the model's strong ability to differentiate between normal and abnormal heartbeats, reinforcing its potential for clinical use in early detection of cardiovascular conditions.

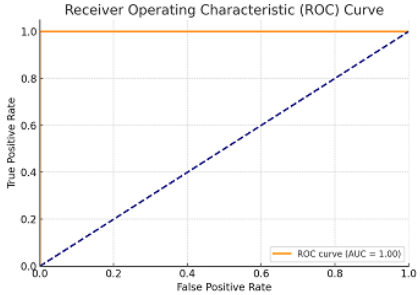


Fig 7. ROC Curve

Confusion Matrix: The confusion matrix for the OSASN-SqueezeNet model highlights its classification performance on ECG heartbeat data. Out of the total predictions, the

model correctly identified 47 abnormal heartbeats (true positives) and 45 normal beats (true negatives), with only 5 false positives and 3 false negatives. The high number of correct classifications emphasizes the model's precision and sensitivity in detecting critical cardiac arrhythmias from ECG signals. Fig 8 displays the confusion matrix values.

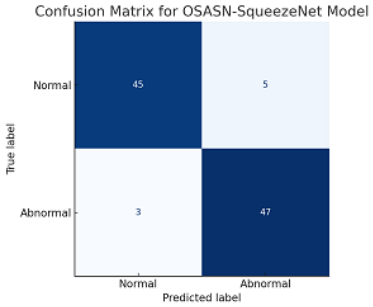


Fig 8. Confusion matrix

7- Comparative Analysis

In the evaluation of different classification methods for heartbeat analysis, various metrics have been employed to assess their performance. The table presents precision, sensitivity, and specificity scores for five distinct methods: K-Nearest Neighbors (KNN), Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM), and the proposed OSASN SqueezeNet. The accuracy of positive predictions is indicated by precision values, and OSASN SqueezeNet had the highest score at 0.95. With a score of 0.94, sensitivity—which stands for the capacity to accurately detect positive instances—also performs better in OSASN SqueezeNet. Specificity, which gauges the capacity to accurately detect negative instances, shows competitive outcomes for all approaches; the suggested OSASN SqueezeNet scored 0.92.

These metrics collectively provide a comprehensive overview of the efficacy of each method in accurately classifying heartbeats, with the proposed OSASN SqueezeNet exhibiting notable strengths in precision and sensitivity.

The table 1 presents performance metrics, such as accuracy, for different classification models including KNN, NN (Neural Network), DT (Decision Tree), SVM (Support Vector Machine), and a proposed model named OSASN SqueezeNet. Each row corresponds to a specific evaluation criterion (e.g., precision, recall, F1-score), and each column represents the respective model.

The provided table presents a comparison of various metrics and methods for evaluating performance. Each row corresponds to a specific metric (Precision, Sensitivity, and Specificity), while each column represents a different method or algorithm (KNN, NN, DT, SVM, and the proposed OSASN SqueezeNet). The numerical values in the table indicate the performance scores or results associated with each combination of metric and method.

Three critical performance metrics—precision, specificity, and sensitivity—are part of the evaluation criteria for classifying heartbeats in ECG signal data. The ratio of true positive cases to the total of true positive and false positive values during the classification process is known as precision, which is a part of accuracy. This metric, which includes all positive values and is expressed as such, is especially pertinent for positive classifications.

The table offers a thorough summary of how well various approaches perform based on three important evaluation criteria. The corresponding method performs better for the specified metric if the numerical value is higher. These accuracy values help determine each model's suitability for the particular classification task at hand by offering insights into how well it performs across a range of evaluation criteria. The proposed OSASN SqueezeNet demonstrates competitive accuracy, suggesting its potential efficacy in the given context.

7-1- Comparative Performance Bar Graph

This comparative bar chart illustrates the performance metrics of five classification models — KNN, Neural Network (NN), Decision Tree (DT), SVM, and the proposed OSASN-SqueezeNet — in classifying ECG heartbeat signals. The OSASN-SqueezeNet model achieved the highest precision (0.93), sensitivity (0.94), and specificity (0.91), outperforming all other baseline classifiers. These results underscore the efficacy of integrating lightweight deep learning with hybrid optimization for reliable heartbeat detection.

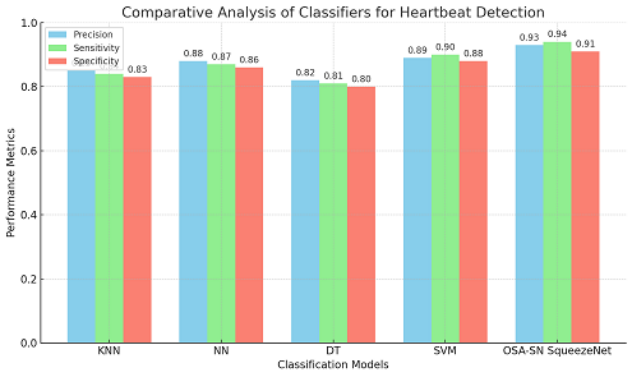


Fig 9. Comparative Analysis

Fig 9 shows the Comparative analysis of classification models in terms of precision, sensitivity, and specificity for ECG heartbeat classification. The proposed OSASN-SqueezeNet model consistently outperformed baseline methods across all evaluation metrics.

Table 1. Comparative Analysis of Classifiers

Classifier	Precision	Sensitivity (Recall)	Specificity
K-Nearest Neighbors (KNN)	0.85	0.84	0.83
Neural Network (NN)	0.88	0.87	0.86
Decision Tree (DT)	0.82	0.81	0.80
Support Vector Machine (SVM)	0.89	0.90	0.88
Proposed OSASN-SqueezeNet	**0.93**	**0.94**	**0.91**

8- Conclusion

The identification of cardiac arrhythmias is crucial for human well-being due to the potential life-threatening nature of these conditions. Among various methods for heartbeat detection, the most widely adopted approach involves the analysis of Electrocardiogram (ECG) signal data. Leveraging data mining techniques proves to be beneficial for uncovering significant patterns, automating tasks, and extracting pertinent records related to heartbeats, aiding in the detection of infectious diseases. In this study, heartbeat classification is conducted using SqueezeNet, trained by a hybridized algorithm.

Initially, the incoming ECG signal data undergoes preprocessing by a median filter, followed by the extraction of two types of characteristics: medical features and statistical features. The Optimization Search Algorithm (OSA-SN) is then used to mine activity patterns from the retrieved features. The final step involves classifying heartbeats using a rule matrix derived from the mined output, employing SqueezeNet trained by OSASN. OSASN is a hybridization of Taylor series and OSA (Optimization Search Algorithm). The performance of this methodology is evaluated using three key metrics: precision, sensitivity, and specificity, with maximum values of 0.91, 0.94, and 0.93, respectively. For future improvements, the technique could benefit from the integration of superior optimization techniques and other advanced deep learning methods to enhance the overall performance metrics.

References

[1] M. Sornalakshmi, S. Balamurali, M. Venkatesulu, M. N. Krishnan, L. K. Ramasamy, S. Kadry, and S. Lim, "An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data," Bulletin of Electrical Engineering and Informatics, vol. 10, no. 1, pp. 390–403, 2021.

[2] X. Chang, C. Zhou, Z. Wang, and P. Hu, "A novel method for mining sequential patterns in datasets," in Sixth International Conference on Intelligent Systems Design and Applications, vol. 1. IEEE, 2006, pp. 611–615.

[3] H. Estiri, S. Vasey, and S. N. Murphy, "Transitive sequential pattern mining for discrete clinical data," in Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18. Springer, 2020, pp. 414–424.

[4] S. B. Golas, T. Shibahara, S. Agboola, H. Otaki, J. Sato, T. Nakae, T. Hisamitsu, G. Kojima, J. Felsted, S. Kakarmath et al., "A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data," BMC medical informatics and decision making, vol. 18, no. 1, pp. 1–17, 2018.

[5] N. Bala, S. N. Price, C. M. Horan, M. W. Gerber, and E. M. Taveras, "Use of telehealth to enhance care in a family-centered childhood obesity intervention," Clinical Pediatrics, vol. 58, no. 7, pp. 789– 797, 2019.

[6] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.

[7] H. Yoo, S. Han, and K. Chung, "A frequency pattern mining model based on deep neural network for real-time classification of heart conditions," in Healthcare, vol. 8, no. 3. MDPI, 2020, p. 234.

[8] J. Jin, W. Sun, F. Al-Turjman, M. B. Khan, and X. Yang, "Activity pattern mining for healthcare," Ieee Access, vol. 8, pp. 56730– 56738, 2020.

[9] S. A. Mangai, B. R. Sankar, and K. Alagarsamy, "Taylor series prediction of time series data with error propagated by artificial neural network," International Journal of Computer Applications, vol. 89, no. 1, 2014.

[10] M. Shanmuga Sundari and V. C. Jadala, "Neurological disease prediction using impaired gait analysis for foot position in cerebellar ataxia by ensemble approach," Automatika, vol. 64, no. 3, pp. 541– 550, 2023.

[11] S. Zhou, J. He, H. Yang, D. Chen, and R. Zhang, "Big data-driven abnormal behavior detection in healthcare based on association rules," IEEE Access, vol. 8, pp. 129002–129011, 2020.

[12] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.

[13] Sornalakshmi, M., Balamurali, S., Venkatesulu, M., Krishnan, M. N., Ramasamy, L. K., Kadry, S., & Lim, S. (2021). An efficient Apriori algorithm for frequent pattern mining using MapReduce in healthcare data. Bulletin of Electrical Engineering and Informatics, 10(1), 390–403.

[14] Yoo, H., Han, S., & Chung, K. (2020). A frequency pattern mining model based on deep neural network for real-time classification of heart conditions. Healthcare, 8(3), 234.

[15] Estiri, H., Vasey, S., & Murphy, S. N. (2020). Transitive sequential pattern mining for discrete clinical data. In Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings (pp. 414–424). Springer.

[16] Jin, J., Sun, W., Al-Turjman, F., Khan, M. B., & Yang, X. (2020). Activity pattern mining for healthcare. IEEE Access, 8, 56730–56738.

[17] Ahmed, S. A., & Nath, B. (2020). Adverse disease agent identification using FP-Growth algorithm in healthcare data

mining. Journal of King Saud University-Computer and Information Sciences. Advance online publication.

[18]L. Penninga, A. K. Lorentzen, and C. Davis, "A telemedicine case series for acute medical emergencies in greenland: a model for austere environments," Telemedicine and e-Health, vol. 26, no. 8, pp. 1066–1070, 2020.

[19]O. N. Oyelade and A. E. Ezugwu, "Ebola optimization search algorithm (eosa): A new metaheuristic algorithm based on the propagation model of ebola virus disease," arXiv preprint arXiv:2106.01416, 2021.

[20]V. C. Jadala et al., "Real-time neurological disease prediction with 3d single pose estimation using mediapipe," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 4s, pp. 595–607, 2024.

[21]V. C. Jadala, S. K. Pasupuleti, P. Yellamma et al., "Deep learning analysis using resnet for early detection of cerebellar ataxia disease," in 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC). IEEE, 2022, pp.

[22]Bindu, V., Sundari, M. S., Pasupuleti, P., Shaik, R., & Pochareddy, J. (2024, November). Revolutionizing Skin Disease Diagnosis: Advanced Image Analysis and Precision Healthcare using EfficientNet. In 2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 604-608). IEEE.–6.

[23]J.-S. Kang, D.-H. Shin, J.-W. Baek, and K. Chung, "Activity recommendation model using rank correlation for chronic stress management," Applied Sciences, vol. 9, no. 20, p. 4284, 2019.

[24]J.-S. Kang, J.-W. Baek, and K. Chung, "Prefixspan based pattern mining using time sliding weight from streaming data," IEEE Access, vol. 8, pp. 124833–124844, 2020.

[25]K. K. Mohbey and S. Kumar, "A parallel approach for high utilitybased frequent pattern mining in a big data environment," Iran Journal of Computer Science, vol. 4, no. 3, pp. 195–200, 2021.

[26] Mariyappan, S. S., Penthala, H. R., Nagaram, A., & Arisham, D. (2024, July). Retina fundus disease gray scale image perception using semantic segmentation model. In AIP Conference Proceedings (Vol. 3028, No. 1). AIP Publishing.

Journal of Information Systems and Telecommunication
Vol.14, No.4, July-September 2026, 225-237
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Soft and Safe Network Slicing with Hierarchical Multi-Agent Reinforcement Learning

Zeinab Sasan[1], Siavash Khorsandi[1*]

[1].Computer Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

Received: 06 Oct 2025 / Revised: 19 Aug 2026 / Accepted: 26 Aug 2026

Abstract

Network slicing has emerged as a key enabler for next-generation mobile networks, allowing multiple virtual network slices to coexist over shared physical infrastructure. Soft network slicing can improve resource utilization by enabling controlled inter-slice resource sharing; however, increased resource sharing may impose a higher burden on donor slices and consequently increase slice dissatisfaction. Achieving efficient resource utilization while controlling slice dissatisfaction and maintaining feasible resource allocation remains challenging under dynamic and heterogeneous traffic demands. In this paper, we propose a Hierarchical Multi-Agent Reinforcement Learning (HMARL) framework for online soft and safe network slicing. The proposed architecture operates at two decision levels: slice-level agents make local admission decisions, while a higher-level meta-controller coordinates inter-slice resource sharing through adaptive borrowing limits. A deterministic safety layer validates the proposed actions before resource allocation and ensures compliance with resource, continuity, and borrowing constraints. The proposed framework is evaluated against learning-based, non-sharing, random, greedy, and heuristic resource-allocation approaches. The training results show that HMARL reaches a stable operating regime with high normalized resource utilization and a low request rejection ratio. Across different scenario classes, HMARL achieves a favorable trade-off between resource utilization and slice dissatisfaction, resulting in competitive overall objective values relative to the considered alternatives. These results indicate that hierarchical coordination combined with deterministic feasibility enforcement provides an effective approach for adaptive resource management in soft network slicing environments.

Keywords: Hierarchical Multi-Agent Reinforcement Learning (HMARL); Soft Network Slicing; Safe Network Slicing; Resource Allocation; Multi-Agent Systems; Safe Reinforment Learning.

1- Introduction

The advent of fifth-generation (5G) and beyond mobile networks has ushered in the era of network slicing, a paradigm that enables multiple logical networks, or slices, to be instantiated on top of a shared physical infrastructure [1]. Each slice is tailored to specific service requirements—such as enhanced mobile broadband, ultra-reliable low-latency communication, or massive machine-type communications—while sharing the same underlying hardware [2], [3]. This flexibility transforms network management but also introduces significant complexity: operators must allocate limited resources among slices dynamically, maintain service-level agreements (SLAs), and maintain an appropriate degree of slice isolation while adapting to heterogeneous and rapidly changing demands [4].

Conventional resource allocation strategies for network slicing typically rely on static provisioning, heuristic admission controls, or optimization models solved offline. While such approaches can be effective in relatively stable settings, they may have difficulty adapting to the variability and unpredictability of dynamic traffic conditions [5], [6]. They may underutilize network capacity, cause unnecessary request rejections, or degrade the quality of service (QoS) of sensitive services [7], [8]. Furthermore, soft and safe network slicing—where controlled resource sharing between slices is permitted while predefined resource-sharing constraints are enforced—poses an even more challenging problem: how to flexibly borrow and lend resources between slices without violating the corresponding resource and borrowing constraints [9], [10], [11].

In this context, reinforcement learning (RL) has emerged as a powerful data-driven approach to sequential decision making under uncertainty [12]. RL allows agents to interact

✉ **Siavash Khorsandi**
khorsandi@aut.ac.ir

with an environment, learn from feedback, and optimize long-term objectives without explicit modeling of the underlying dynamics. Extending RL to multiple cooperating entities leads to multi-agent reinforcement learning (MARL), which is naturally suited to distributed resource allocation problems like network slicing [13]. However, conventional flat MARL architectures—where all agents operate at the same decision layer—may face coordination challenges as the number of slices, actions, or resource constraints increases. They may also converge slowly or produce decisions that require additional constraint handling in dynamic resource allocation environments [14].

To address these challenges, this paper proposes a Hierarchical Multi-Agent Reinforcement Learning (HMARL) framework for online resource management in soft and safe network slicing. The proposed framework decomposes the decision-making process into two hierarchical levels. At the lower level, slice-specific agents evaluate eligible user requests based on their local observations and make admission decisions for individual requests. At the higher level, a meta-controller observes the system-level state and adaptively determines borrowing limits that regulate resource sharing between neighboring slices. This hierarchical structure combines local admission decisions with system-level coordination, allowing the resource-sharing policy to adapt to changing traffic conditions. Before execution, a deterministic safety layer validates the proposed decisions and performs resource allocation while enforcing the predefined resource, borrowing, and continuity constraints.

We evaluate the proposed HMARL framework against several representative alternatives, including HSS, a hierarchical reinforcement-learning baseline; a variant without the meta-controller (w/o Meta); No-Sharing; Random Admission; a greedy heuristic; and HRASS. The evaluation focuses on four key performance metrics: the overall objective value, normalized resource utilization, normalized slice dissatisfaction, and request rejection ratio. In addition, the training behavior of the considered learning-based approaches is examined to assess convergence over the training episodes. The results show that HMARL achieves high resource utilization and a low rejection ratio while maintaining competitive objective values across different scenario classes. Overall, the proposed hierarchical approach provides a favorable trade-off between efficient resource utilization, controlled inter-slice dissatisfaction, and request admission performance.

The contributions of this paper can be summarized as follows:

(1) We formulate online soft network slicing as a hierarchical multi-agent decision-making problem, where slice-level agents generate local admission proposals and a meta-controller coordinates inter-slice resource sharing through adaptive borrowing limits;
(2) We develop an HMARL framework that combines on-policy learning with a deterministic safety layer that validates the proposed decisions and enforces resource, borrowing, and continuity constraints before allocation;
(3) We evaluate the proposed framework against multiple learning-based, heuristic, random, and non-sharing alternatives under heterogeneous traffic scenarios, showing that HMARL achieves high resource utilization, low request rejection, and a favorable utilization–dissatisfaction trade-off.

The remainder of the paper is organized as follows. Section II reviews related work on soft network slicing and reinforcement learning–based resource management. Section III introduces the system model and problem formulation. Section IV describes the proposed HMARL framework together with the baseline methods. Section V presents the experimental setup and evaluation metrics. Section VI discusses the performance results. Finally, Section VII concludes the paper and outlines future research directions.

2- Related Works

In this section, we review the most relevant studies that have addressed network slicing from different perspectives. We first examine the literature on traditional and soft network slicing, highlighting approaches that aim to dynamically allocate resources among multiple virtual slices. We then discuss prior work that leverages multi-agent reinforcement learning (MARL) techniques to manage slice admission and resource sharing in a distributed manner. Finally, we focus on hierarchical MARL frameworks, which introduce additional layers of decision-making to improve coordination and adaptability. This overview positions our proposed HMARL framework within the context of existing research and clarifies how it advances the state of the art.

Sasan et al. [15] proposed the Heuristic Resource Allocation for Soft Slicing (HRASS) framework to overcome the inefficiencies of hard network slicing. Their work formulates soft slicing as a Mixed Integer Linear Programming (MILP) problem with the dual objective of maximizing resource utilization and minimizing slice dissatisfaction. To address the NP-hardness of the problem, the authors develop a heuristic algorithm that achieves near-optimal performance with significantly reduced computational complexity. Experimental results in their study demonstrate that HRASS effectively improves resource utilization while mitigating the limitations of hard slicing.

Li et al. [16] proposed a hierarchical soft-slicing framework for software-defined radio access networks supporting differentiated services with diverse QoS requirements. Their approach splits slicing into two levels: at the network

level, radio resources are pre-allocated to each gNodeB over a large time scale, while at the gNodeB level, these resources are dynamically scheduled at mini-slot granularity to handle short-term request variations. The framework enables time-varying traffic loads of differentiated services across multiple gNodeBs and allows dynamic inter-gNodeB resource sharing to increase multiplexing gains. A case study in their article demonstrates the effectiveness of the approach and outlines open research challenges.

Shan Zhang et al. [9] proposed a service-oriented hierarchical soft slicing framework for cache-enabled vehicular networks to meet diverse, multi-dimensional QoS requirements. In their approach, each slice supports a specific service with logical resource isolation while opportunistically reusing resources to exploit multiplexing gains. The framework is analysed for two representative on-road content services — a time-critical context information service (CIS) and a bandwidth-intensive infotainment service (IS) — with intra- and inter-slice resource reuse. Analytical and simulation results show that the proposed method can increase IS throughput by about 30 % while maintaining the same level of CIS content freshness and service delay.

Shi et al. [17] proposed a two-level soft slicing scheme for 5G-and-beyond radio access networks to simultaneously support ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) services. At the network level, they determine the required radio resources for eMBB and analyse the delay-violation probability for URLLC, formulating an integer non-linear program for network-level preallocation. Given the NP-completeness of the problem, they develop a low-complexity heuristic to obtain near-optimal solutions, and at the gNB level they design a scheduling scheme enabling real-time resource sharing among URLLC services. Simulation results show that their soft-slicing approach satisfies stringent QoS requirements for both service types while achieving significantly higher resource-utilization efficiency than conventional hard slicing schemes.

Jingfang Ding et al. [18] proposed a dynamic switching soft slicing strategy for industrial mixed traffic in 5G networks. Their approach targets two traffic types — periodic delay-sensitive (PDS) and sporadic delay-tolerant (SDT) — and combines a traffic-QoS-aware soft slicing (TQASS) scheme with a resource-efficiency-aware soft slicing (REASS) scheme. The strategy guarantees the reliability of PDS traffic under strict delay constraints while dynamically allocating remaining resources to SDT traffic. Simulation results demonstrate that this dynamic soft slicing approach outperforms existing methods in satisfying the stringent QoS requirements of industrial mixed traffic.

Kaytaz et al. [19] addressed the RAN slicing problem in 5G vehicle-to-everything (V2X) communications and proposed an h-DQN-based Soft Slicing (HSS) method for model-free opportunistic slice management. HSS implements a hierarchical reinforcement learning framework in which a high-level meta-controller selects subgoals and low-level controllers choose actions conditioned on these subgoals and the system state. By decomposing performance constraints and applying deep reinforcement learning at different temporal scales, their approach improves sample efficiency and, according to their results, outperforms traditional and RL-based V2X RAN slice-management schemes in terms of Age of Information (AoI), service delay, and network utility maximization.

Yan Shao et al. [20] formulated inter-slice resource management in dense cellular networks as a multi-agent reinforcement learning problem and enhanced cooperation between base-station agents using graph attention networks (GAT), achieving superior performance over existing methods in extensive simulations. Nassima Toumi et al. [21] addressed the multi-domain service function chaining (SFC) placement problem by leveraging deep reinforcement learning. They proposed a hierarchical architecture with local and multi-domain agents trained on different DRL models to perform SFC and sub-SFC placement while meeting SLA requirements.

Xianfu Chen et al. [22] tackled two-timescale radio resource management in O-RAN slicing by proposing a meta Markov decision process formulation. They decomposed the problem into a long-timescale master and a short-timescale subproblem, solved via a hierarchical reinforcement learning mechanism combining deep RL and meta-RL. Their results showed superior performance over state-of-the-art baselines. Hasan Anıl Akyıldız et al. [23] proposed a multi-agent hierarchical reinforcement learning approach for joint radio resource allocation between URLLC and eMBB slices under random traffic arrivals. A main agent performs inter-slice allocation, while URLLC and eMBB sub-agents handle intra-slice allocation. The state space incorporates queue occupancy and channel quality, and actions correspond to user-level resource assignments. Their results show sub-millisecond URLLC delays with minimal eMBB throughput degradation.

Hnin Pann Phyu et al. [24] addressed the energy–QoS trade-off in sliced mobile networks. They formulated a joint slice activation/deactivation and user association problem aimed at minimizing energy consumption while maintaining QoS. Their proposed multi-agent cooperative framework (ICE-CREAM) uses agents at two granularity levels that share information with neighbors and are trained on a global reward. Evaluation on a real-world dataset of French mobile service consumption showed more than 30% energy efficiency improvement over always-active slices while preserving QoS. Tianlun Hu et al. [25] proposed a multi-agent deep reinforcement learning approach for dynamic, slice-aware radio resource partitioning across multiple cells. Their framework tackles inter-cell interference and inter-

slice constraints by introducing two distributed coordination schemes that allow agents to cooperate. System-level simulations demonstrate that the proposed method achieves over two-fold improvement in resource efficiency and better delay performance compared with centralized baselines.

Guorong Zhou et al. [26] addressed the challenge of multi-objective optimization in RAN slicing, where throughput, delay, and coverage of three slice types (high-bandwidth, low-delay, wide-coverage) must be jointly optimized over a shared infrastructure. They proposed an improved multi-agent deep deterministic policy gradient (IMADDPG) algorithm with centralized training and distributed execution to solve the NP-hard problem. Using a rank voting mechanism to approach near-Pareto optimal solutions, their method outperformed scalar-utility and benchmark algorithms and achieved a flexible tradeoff among competing slice objectives.

Federico Mason et al. [14] investigated dynamic resource allocation in network slicing using a distributed Deep Reinforcement Learning framework. Their approach deploys multiple cooperating agents trained with the Advantage Actor-Critic (A2C) algorithm to handle continuous action spaces. Through extensive simulations, they showed that the proposed DRL-based strategy outperforms efficient empirical algorithms, achieving higher adaptability to varying scenarios without requiring additional training. Muhammad Sulaiman et al. [13] addressed joint network slicing and slice admission control in 5G Cloud-RANs by proposing a multi-agent Deep Reinforcement Learning (DRL) approach. Their framework learns the dynamics of slice-request traffic and jointly optimizes slice placement and admission decisions. Compared with heuristic baselines and single-agent DRL, the proposed multi-agent DRL achieves up to 18% and 3.8% higher long-term revenue, respectively, and exhibits improved robustness when generalized to unseen scenarios.

Yohan Kim et al. [12] proposed an RL-based dynamic resource allocation framework for end-to-end network slicing in hierarchical MEC environments. They modelled the problem as a Markov Decision Process (MDP) and used proximal policy optimization (PPO) to develop independently-collaborative and jointly-collaborative allocation algorithms. Their results show that the proposed methods can effectively recognize slice request characteristics and allocate resources efficiently while maintaining a high QoS satisfaction rate for heterogeneous 5G slices.

Hao Zhou et al. [27] developed a multi-agent reinforcement learning (MARL) approach for radio resource slicing in 5G networks. In their framework, each slice acts as an intelligent agent competing for limited radio resources, and a correlated Q-learning scheme (CO-QRA) is applied for inter-slice resource block allocation. They compare CO-QRA with Nash Q-learning (NQL), Latency-Reliability-Throughput Q-learning (LRTQ), and priority proportional fairness (PPF). Simulation results show that CO-QRA significantly reduces latency and packet drop rate while improving throughput compared with these baselines.

Despite the growing body of work on network slicing, the joint consideration of soft inter-slice resource sharing and hierarchical multi-agent decision making remains relatively limited. Existing soft-slicing approaches mainly rely on optimization, heuristic, or application-specific hierarchical mechanisms, while many MARL-based slicing methods focus on resource allocation without explicitly incorporating controlled resource sharing and slice dissatisfaction into the decision process. Moreover, ensuring that learning-based admission decisions remain feasible under resource, borrowing, and continuity constraints requires an explicit constraint-handling mechanism. To address these gaps, we propose an online hierarchical multi-agent framework in which slice-level agents perform admission decisions, a meta-controller coordinates inter-slice borrowing, and a deterministic safety layer validates and executes feasible resource allocations.

3- System Model and Problem Formulation

We consider a soft network slicing environment in which multiple logical slices share the available radio resources, with each slice assigned a predefined base resource share. User requests arrive dynamically with heterogeneous resource demands and service requirements. When its own available resources are insufficient, a slice may temporarily borrow unused resources from eligible neighboring slices, subject to predefined borrowing limits. While such controlled sharing can improve resource utilization, lending resources imposes a burden on donor slices and may increase slice dissatisfaction. Therefore, resource sharing must be adaptively controlled to balance resource utilization and slice dissatisfaction while maintaining feasible allocations.

3-1- System Entities

Slices: We denote the set of network slices by $\mathcal{M} = \{1, 2, \ldots, M\}$. Each slice m corresponds to a specific type of service or tenant with its own arrival processes, deadlines, and resource demands.

Jobs/Requests: Each incoming job for slice m is characterized by an arrival time a_i, a processing duration τ_i, a resource demand r_i, and a deadline d_i. In addition, a sensitivity indicator is included as part of the request information available to the corresponding slice-level agent when making admission decisions.

Resources: The system has a total resource capacity *R*, shared among all slices over discrete time slots. Each slice can primarily use its allocated base quota, while unused

resources may be temporarily borrowed by eligible neighboring slices. Inter-slice borrowing is permitted only between slices connected in the predefined sharing topology and is limited by the borrowing caps determined by the meta-controller.

Safety Layer: A deterministic safety layer validates admission decisions before resource allocation. For each accepted request, it first checks the available capacity of the request's own slice. If additional resources are required, borrowing is considered only from eligible neighboring slices and is constrained by the donor's available headroom and the remaining borrowing cap. The allocation is committed only when all resource and continuity constraints can be satisfied.

3-2- Decision Structure

At each time step t, the system must decide:

1. **Demission Control:** For each slice, the corresponding lower-level agent determines which eligible user requests should be admitted based on the current request and system information.
2. **Meta-Control**: At a slower decision timescale, the meta-controller determines the inter-slice borrowing caps that regulate how much resource sharing is permitted between neighboring slices.

These decisions form a two-level hierarchical control structure in which lower-level slice agents perform local admission decisions, while the meta-controller coordinates inter-slice resource sharing. The resulting decisions are subsequently validated and executed by the deterministic safety layer, which enforces the resource, borrowing, and continuity constraints.

3-3- Objective

The system-level objective is to efficiently utilize the available radio resources while controlling slice dissatisfaction arising from inter-slice resource sharing. At each time slot t, the instantaneous resource utilization U_t is defined as the ratio of the total amount of allocated resources to the overall available resource capacity:

$$U_t = \frac{R_t^{\text{used}}}{R} \tag{1}$$

Where R_t^{used} denotes the total amount of resources in use at time slot t, and R is the total available resource capacity.

To quantify the cost associated with inter-slice resource sharing, we define slice dissatisfaction according to the fraction of a slice's base resource quota that is temporarily lent to other eligible slices. Let $L_m(t)$ denote the amount of resources lent by slice m at time slot t, and let q_m denote its predefined base quota. The instantaneous system-wide dissatisfaction D_t is defined as:

$$D_t = \frac{1}{M}\sum_{m=1}^{M}\frac{L_m(t)}{q_m} \tag{2}$$

A value of D_t close to zero indicates little or no inter-slice sharing, whereas a larger value represents a higher resource-sharing burden on donor slices.

Over an episode consisting of T time slots, the average resource utilization U is calculated as:

$$U = \frac{1}{T}\sum_{t=1}^{T}U_t \tag{3}$$

while the average slice dissatisfaction D is given by:

$$D = \frac{1}{T}\sum_{t=1}^{T}D_t \tag{4}$$

For performance evaluation, these two metrics are combined into the following system-level objective:

$$J_{\text{eval}} = \alpha U - \beta D \tag{5}$$

where α and β are non-negative weighting coefficients representing the relative importance of resource utilization and slice dissatisfaction, respectively.

3-4- Problem Formulation

The resulting problem can be formulated as a hierarchical multi-agent decision-making task. At the lower level, each slice-specific agent evaluates eligible user requests using a fixed-dimensional observation that captures the request slack, resource demand, processing duration, sensitivity indicator, and the current system-level acceptance ratio, and then makes an admission decision for each request. At the higher level, a meta-controller operates at a slower timescale and observes compact system-level information for each slice, including its base resource share, pending-request ratio, accepted-request ratio, and remaining borrowing-cap ratio. Based on this information, the meta-controller adaptively determines the inter-slice borrowing caps that regulate resource sharing between eligible neighboring slices. Before execution, the proposed admission decisions are processed by a deterministic safety layer that checks resource availability, donor headroom, borrowing limits, and continuity requirements before committing an allocation.

This hierarchical formulation separates fine-grained admission decisions from system-level coordination of inter-slice sharing. By combining adaptive learning with deterministic feasibility enforcement, it provides a suitable foundation for achieving efficient resource utilization while controlling the cost of inter-slice borrowing. The resulting HMARL framework is described in the next section.

4- Proposed HMARL Framework

4-1- Hierarchical Architecture and Decision Process

The proposed Hierarchical Multi-Agent Reinforcement Learning (HMARL) framework is designed for online resource management in a soft network slicing environment. The framework combines decentralized admission decisions at the slice level with system-wide coordination of inter-slice resource sharing. It consists of three main components: a set of low-level slice actors, a higher-level meta-controller, and a deterministic safety layer.

At the lower level, each network slice m is associated with an independent actor π_{θ_m}. The slice-level actor evaluates the currently eligible requests belonging to its own slice and determines whether each request should be proposed for admission. These actors operate at every time slot and focus on fine-grained admission control based on request-level and system-level information. Their outputs are admission proposals that are subsequently processed by the resource-allocation mechanism.

At the higher level, a meta-controller π_{θ_H} operates at a slower time scale, once every K time slots. The meta-controller observes a compact summary of the state of all slices and determines the inter-slice borrowing caps that regulate resource sharing between eligible neighboring slices. The predefined base resource quota of each slice remains unchanged during operation; therefore, the role of the meta-controller is limited to adapting the allowable degree of inter-slice borrowing according to the current network conditions.

The decisions generated by the learning policies are not executed directly. Instead, the proposed admission decisions and the borrowing limits determined by the meta-controller are provided to a deterministic safety layer. The safety layer verifies resource availability and checks the relevant borrowing, topology, and continuity constraints before committing an allocation. If the requirements of a proposed admission cannot be satisfied, the request is not allocated even when the corresponding low-level actor has proposed its admission.

This hierarchical design separates local admission control from global coordination of inter-slice resource sharing. The low-level actors can therefore focus on individual request decisions, while the meta-controller regulates the overall sharing policy at a slower time scale. At the same time, deterministic feasibility checking prevents learning-based actions from directly producing resource allocations that violate the predefined system constraints.

4-2- State and Action Spaces

Low-Level Slice Agents

Each network slice *m* is associated with a low-level actor that makes admission decisions for the requests belonging to that slice. To accommodate a variable number of eligible requests without changing the neural-network dimensions, each request is evaluated independently using a fixed-dimensional observation vector.

For an eligible request i of slice m at time slot t, the local observation is defined as

$$o_{m,i,t} = [\hat{s}_{i,t}, \hat{r}_i, \hat{\tau}_i, z_i, \rho_t^{acc}] \qquad (6)$$

where $\hat{s}_{i,t}$ denotes the remaining slack of request i, $\hat{r}_i$ is its resource demand, $\hat{\tau}_i$ represents its service duration, $z_i \in 0,1$ is the sensitivity indicator, and $\rho_t^{acc} \in [0,1]$ denotes the current system-level acceptance ratio. Thus, the low-level actor receives a five-dimensional observation for each eligible request. The slice actor π_{θ_m} maps the observation vector to an admission probability:

$$p_{m,i,t} = \pi_{\theta_m}(o_{m,i,t}) \qquad (7)$$

During training, the corresponding binary action is sampled from a Bernoulli distribution:

$$a_{m,i,t} \sim \text{Bernoulli}(p_{m,i,t}) \qquad (8)$$

With

$$a_{m,i,t} \in 0,1. \qquad (9)$$

Here, $a_{\{m,i,t\}} = 1$ indicates that the low-level policy proposes request i for admission, whereas $a_{\{m,i,t\}} = 0$ indicates that the request is not proposed for admission at the current time slot. The admission proposal does not itself guarantee resource allocation, since every proposed request must subsequently pass the deterministic feasibility checks performed by the safety layer.

This per-request formulation avoids constructing variable-length state and action vectors. Regardless of the number of eligible users at a given time slot, the same five-dimensional actor input and one-dimensional Bernoulli output are applied independently to each request.

Meta-Controller

The meta-controller operates at a slower time scale and makes one decision every *K* time slots. At each meta-decision instant, it observes a compact system-level representation that summarizes the condition of all *M* slices. For each slice *m*, the high-level observation vector is defined as

$$x_m^H = [\hat{q}_m, \rho_m^{pend}, \rho_m^{acc}, \rho_m^{cap}] \qquad (10)$$

where $\hat{q}_m$ denotes the base resource share of slice *m*, $\rho_m^{pend} \in [0,1]$ is its pending-request ratio, $\rho_m^{acc} \in [0,1]$ is its accepted-request ratio, and $\rho_m^{cap} \in [0,1]$ represents its remaining borrowing-cap ratio. The complete high-level

state is therefore represented by an $M \times 4$ matrix $X_t^H \in \mathbb{R}^{M \times 4}$.

Based on X_t^H , the meta-controller generates continuous latent actions using a parameterized Normal policy. These actions are transformed into the inter-slice borrowing-cap matrix

$$C_t = [c_{m \leftarrow n}(t)] \in \mathbb{R}^{M \times M}, \qquad (11)$$

where $c_{m \leftarrow n}(t)$ represents the maximum amount of resources that requesting slice *m* is permitted to borrow from donor slice *n* during the corresponding meta-control interval.

The meta-controller therefore controls only the allowable degree of inter-slice borrowing. The resulting borrowing caps are subsequently used by the deterministic safety layer when validating and executing the low-level admission proposals.

4-3- Deterministic Safety Layer

The admission decisions generated by the low-level actors are treated as admission proposals and are not directly translated into resource allocations. Before execution, all proposed requests are processed by a deterministic safety layer that verifies their feasibility with respect to the available resources, slice quotas, borrowing limits, sharing topology, and service requirements. Therefore, the learning policies determine which requests should be considered for admission, while the safety layer determines whether these proposals can actually be executed.

At each time slot *t*, the safety layer first collects all requests proposed for admission by the low-level actors. Since several requests may be proposed simultaneously, they are processed according to a deterministic priority rule. For each request *i*, the remaining slack is computed as

$$s_{i,t} = d_i - (t + \tau_i - 1), \qquad (12)$$

where d_i denotes the request deadline and τ_i is its required service duration. Requests with smaller slack are processed first because they have less scheduling flexibility. If two or more requests have the same slack, requests with larger resource demands are given higher priority, followed by those with shorter service durations. This ordering ensures that simultaneous admission proposals are handled consistently rather than being allocated in an arbitrary order.

For each request in the resulting priority order, the safety layer first checks whether the required resources can be supported over the complete service interval from time slot t to $t + \tau_i - 1$. It then attempts to satisfy the resource demand using the available capacity associated with the request's own slice. If sufficient local capacity is available throughout the required service interval, the request can be admitted without inter-slice borrowing.

When the available capacity of the requesting slice is insufficient, the remaining resource deficit is considered for inter-slice borrowing. Only slices connected to the requesting slice through the predefined sharing topology are considered as potential donors. For each eligible donor, the safety layer determines the amount of resource that can be provided according to two conditions: the donor must have sufficient available headroom, and the corresponding remaining borrowing cap must permit the requested transfer. Hence, the amount borrowed from a donor cannot exceed either its available resource headroom or the remaining cap associated with that slice pair $c_{m \leftarrow n}(t)$.

If multiple eligible donors are available, the safety layer examines the admissible borrowing opportunities until the remaining demand is satisfied or no additional feasible resource can be obtained. The request is admitted only when its complete resource demand can be supported over the entire required service interval. Partial allocations are not committed. If the complete demand cannot be satisfied using the combination of local and admissibly borrowed resources, the proposed admission is rejected and no tentative resource assignment is retained.

After a request is successfully admitted, the corresponding resource occupancy and remaining borrowing capacities are immediately updated. The next request in the priority order is therefore evaluated according to the updated system state. This sequential update is particularly important when several requests are proposed in the same time slot, since resources committed to an earlier request are no longer considered available to subsequent requests.

In addition, the safety layer prevents conflicting assignments by ensuring that the same physical resource is not simultaneously allocated to multiple requests in a given time slot. Resource availability is checked across the entire service duration of each admitted request, thereby preserving the required allocation for the requested processing interval.

Overall, the deterministic safety layer acts as a feasibility-enforcement mechanism between the learned policies and the actual resource allocator. It does not introduce an additional learned policy and does not modify the objectives of the low-level actors or the meta-controller. Instead, it ensures that only admission proposals satisfying the explicitly defined resource, borrowing, topology, and service-continuity constraints are executed. Thus, feasibility of the executed allocations is enforced deterministically, without assuming that the reinforcement-learning policies themselves learn or guarantee constraint satisfaction.

4-4- Reward Design and Advantage Estimation

The reward function used to train the HMARL policies is designed to encourage efficient resource utilization while discouraging excessive dependence on inter-slice

borrowing. At each time slot *t*, the training reward is defined as

$$r_t = \alpha_r \tilde{U}_t - \beta_r \tilde{D}_t + \eta_r B_t^{direct} - \xi_r P_t^{borrow}, \quad (13)$$

where $\tilde{U}_t$ denotes the normalized resource-utilization component and $\tilde{D}_t$ denotes the normalized slice-dissatisfaction component used during training. The term B_t^{direct} provides an additional positive signal when requests are served without requiring inter-slice borrowing, while P_t^{borrow} penalizes excessive use of borrowed resources. The coefficients α_r , β_r , η_r , and ξ_r control the relative contributions of these components.

4-5- PPO-Based Policy Optimization

The proposed framework follows centralized training with hierarchical execution and local admission decisions at the lower level. Two independent state-value critics are used during training. A shared centralized critic is used for the low-level slice actors, while a separate critic is associated with the high-level meta-controller. The two critics have independent parameters, although both operate on a compact global representation of the system state.

Let $V_{\phi_L}(g_t)$ denote the value function of the shared low-level critic, and let $V_{\phi_H}(g_t)$ denote the value function of the high-level critic, where $g_t \in \mathbb{R}^6$ represents the compact global state supplied to the critics. The critics are used only during training to estimate expected future returns and construct the advantage signals required for policy optimization.

For the low-level policies, the temporal-difference residual at time slot *t* is

$$\delta_t^L = r_t + \gamma V_{\phi_L}(g_{t+1}) - V_{\phi_L}(g_t), \quad (14)$$

where $\gamma \in [0,1]$ is the discount factor.

The corresponding low-level advantage is estimated using Generalized Advantage Estimation (GAE):

$$\hat{A}_t^L = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}^L, \quad (15)$$

where $\lambda \in [0,1]$ is the GAE parameter. The parameter λ controls the trade-off between the bias and variance of the advantage estimate, while γ determines the relative importance of future rewards.

Because the meta-controller operates at a slower time scale, a high-level learning sample corresponds to the interval between two consecutive meta-decisions rather than to a single time slot. Let t_j denote the time of the j-th meta-decision and let Δ_j denote the number of time slots until the next meta-decision. The reward associated with this high-level transition is obtained by aggregating the low-level rewards over the corresponding interval:

$$R_j^H = \sum_{l=0}^{\Delta_j - 1} \gamma^l r_{t_j + l}. \quad (16)$$

the high-level temporal-difference residual can then be expressed as

$$\delta_j^H = R_j^H + \gamma^{\Delta_j} V_{\phi_H}(g_{t_{j+1}}) - V_{\phi_H}(g_{t_j}). \quad (17)$$

The high-level advantage $\hat{A}_j^H$ is subsequently obtained using the same GAE principle over the sequence of meta-level transitions. In this way, the low-level actors receive advantage signals corresponding to fine-grained admission decisions, whereas the meta-controller receives feedback that reflects the cumulative effect of its borrowing-cap decisions over a longer control interval.

To improve numerical stability during policy optimization, the learning signals used for advantage estimation are normalized during training. This normalization is applied only to the learning procedure and does not change the definitions of the performance metrics reported in the evaluation results.

The policies at both levels of the proposed framework are trained using Proximal Policy Optimization (PPO). For a policy with parameters θ, the probability ratio between the current policy and the policy used to collect the training samples is defined as

$$\rho_t(\theta) = \frac{\pi_\theta(a_t \mid o_t)}{\pi_{\theta^{old}}(a_t \mid o_t)}. \quad (18)$$

PPO updates the policy using the clipped surrogate objective

$$L^{clip}(\theta) = \mathbb{E}_t\left[\min\left(\rho_t(\theta)\hat{A}_t, \text{clip}\left(\rho_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right], \quad (19)$$

where $\hat{A}_t$ is the advantage estimate obtained using GAE and ϵ is the PPO clipping parameter. The clipping operation limits excessive changes between successive policy updates and improves the stability of on-policy training.

The same PPO principle is applied separately at the two hierarchical levels. For the low-level actors, the policy corresponds to the admission decisions generated for individual requests, and the updates are based on the low-level advantages computed at the time-slot scale. For the meta-controller, the policy corresponds to the continuous actions used to determine the borrowing caps, and its updates are performed using the high-level advantages computed at the slower meta-decision scale.

5- Experimental Setup and Evaluation Metrics

5-1- Simulation Environment and Traffic Scenarios

The proposed framework is evaluated using a Python/PyTorch-based simulator developed for online soft network slicing. Multiple logical network slices share the available resources, while controlled inter-slice borrowing is permitted only between eligible neighboring slices according to a predefined sharing topology. Each slice is assigned a fixed base resource share, and temporary resource borrowing is regulated by the corresponding inter-slice borrowing caps.

At the beginning of each episode, a new set of user requests is generated for each slice. Each request is characterized by an arrival time, resource demand, service duration, deadline, and a binary sensitivity indicator. These request characteristics vary across users and episodes to represent heterogeneous and time-varying traffic conditions.

To evaluate the methods under different levels of scheduling difficulty, five traffic scenario classes are considered. The scenario classes differ in the temporal distribution of request arrivals and the tightness of their scheduling windows. They range from relatively dispersed traffic with greater scheduling flexibility to more concentrated traffic with tighter timing requirements. Accordingly, the considered scenarios expose the evaluated methods to different levels of temporal contention and resource-allocation difficulty.

For each evaluation instance, a fresh traffic realization is generated according to the corresponding scenario class. The performance of the considered methods is subsequently evaluated across all scenario classes to provide a representative assessment under heterogeneous traffic conditions.

5-2- Training and Simulation Configuration

The main simulation and learning parameters are summarized in Table I. The same environmental settings and traffic-generation rules are used when evaluating all methods to provide a consistent comparison. For learning-based methods, the policy networks are independently initialized and trained using fresh training instances. Unless otherwise stated, the parameter values listed in Table I are kept fixed throughout the experiments.

Table 1: Simulation and Training Parameters

Parameter	Value
Number of slices (M)	5
Total resource capacity (R)	40
Episode horizon (T)	30 slots
Base resource proportions	[0.2, 0.2, 0.4, 0.1, 0.1]
Requests per slice	7
Meta-controller interval (K)	5 slots
Training episodes	2000
Learning rate	3×10^{-4}
Discount factor (γ)	0.99
GAE parameter (λ)	0.95
PPO clipping parameter (ε)	0.30
PPO optimization epochs	4
Evaluation-objective weights	$\alpha = 0.6$, $\beta = 0.4$

5-3- Baseline and Comparison Methods

To evaluate the proposed HMARL framework, we compare it with several learning-based, heuristic, and non-sharing approaches. All methods are evaluated under the same traffic scenarios and environmental settings to provide a consistent comparison.

h-DQN based Soft Slicing (HSS) is included as a hierarchical reinforcement-learning baseline [19]. In this approach, a high-level controller selects a slice as the current subgoal, while a lower-level controller performs the corresponding admission or allocation decisions. Unlike the proposed HMARL framework, HSS follows a sequential subgoal-based hierarchy rather than jointly coordinating all slices through adaptive inter-slice borrowing caps.

The w/o Meta variant removes the meta-controller from HMARL while retaining the low-level slice actors. Consequently, admission decisions are still learned, but inter-slice sharing is not adaptively coordinated at the higher level. This variant is used to assess the contribution of hierarchical meta-level coordination.

No-Sharing represents a hard-isolation configuration in which inter-slice borrowing is disabled by setting the borrowing caps to zero. Each slice can therefore use only its predefined resource share. This baseline illustrates the performance obtained when resource isolation is preserved at the expense of flexible inter-slice sharing.

Random Admission provides a simple non-learning reference. Eligible requests are proposed for admission according to a fixed Bernoulli probability, while the feasibility of the resulting proposals is handled by the same resource allocation constraints used in the simulation environment.

The Greedy method provides a lightweight heuristic reference. Requests are prioritized according to their scheduling urgency and resource requirements, and feasible requests are admitted sequentially whenever sufficient

resources are available. Unlike the learning-based approaches, Greedy does not learn a policy from interaction with the environment.

Heuristic Resource Allocation for Soft Slicing (HRASS) is included as a heuristic resource-allocation reference that explicitly searches for feasible assignments while considering request timing, resource requirements, and admissible inter-slice resource sharing [15]. It uses broader request information to construct complete allocations and therefore provides a useful optimization-oriented reference for assessing the performance of the online learning approaches.

Together, these comparison methods cover different levels of decision-making complexity, ranging from random and isolated admission to heuristic and hierarchical learning-based resource management.

5-4- Evaluation Metrics

The performance of the considered methods is evaluated using four main metrics: the overall reward value, resource utilization, slice dissatisfaction, and request rejection ratio. Resource utilization and slice dissatisfaction are computed according to the definitions introduced in Section 3. Their normalized values are reported in the comparative results to facilitate consistent comparison across different traffic scenarios.

The request rejection ratio measures the fraction of generated requests that are not successfully admitted and allocated during an episode. A lower rejection ratio therefore indicates that a larger proportion of the offered traffic can be successfully served. These metrics are considered jointly, since a method may achieve low dissatisfaction by limiting inter-slice borrowing while simultaneously experiencing lower resource utilization or a higher request rejection ratio.

6- Performance Evaluation

This section presents and discusses the performance results of the proposed HMARL framework. The evaluation is conducted in two stages. First, the training behavior of HMARL is examined in terms of the evaluation objective, normalized resource utilization, normalized slice dissatisfaction, and request rejection ratio. Then, the performance of HMARL is compared with the considered learning-based, heuristic, and non-sharing methods across different traffic scenario classes using the same evaluation metrics.

6-1- Training Convergence

Fig. 1 illustrates the evolution of the main performance metrics over 2000 training episodes. The figure reports the training reward, normalized resource utilization, normalized slice dissatisfaction, and request rejection ratio. The trajectories show that HMARL undergoes a relatively short initial adaptation phase and subsequently reaches a stable operating regime, with only moderate fluctuations over the remaining training episodes.

As shown in Fig. 1(a), the reward value of HMARL increases rapidly during the early stage of training and stabilizes at approximately 0.23 after the initial adaptation period. It remains consistently above the other considered methods throughout most of the training process. This behavior indicates that the hierarchical policy progressively learns decisions that provide a favorable balance between resource utilization and the cost associated with inter-slice sharing. The relatively small fluctuations observed after stabilization suggest consistent performance under dynamically generated traffic instances.

The utilization trend in Fig. 1(b) provides further insight into this behavior. HMARL rapidly increases its normalized resource utilization to approximately 0.70 and maintains this level for most of the training process. The remaining methods stabilize at noticeably lower utilization levels. This result suggests that hierarchical coordination and adaptive borrowing allow HMARL to make more effective use of the available radio resources.

The higher utilization achieved by HMARL is accompanied by higher normalized slice dissatisfaction, as illustrated in Fig. 1(c). After the initial adaptation phase, the dissatisfaction of HMARL stabilizes at approximately 0.46–0.49. This behavior is associated with more active use of controlled inter-slice borrowing, which imposes a greater sharing burden on donor slices. In contrast, No-Sharing maintains a dissatisfaction value close to zero because inter-slice borrowing is disabled by construction.

Fig. 1(d) shows a complementary trend in the request rejection ratio. HMARL exhibits a sharp reduction in rejection during the early training stage, after which the ratio stabilizes at approximately 0.19. This is lower than the corresponding levels observed for the other methods in the figure. The simultaneous increase in utilization and reduction in rejection indicate that HMARL learns to serve a larger fraction of the offered requests while making greater use of controlled inter-slice sharing.

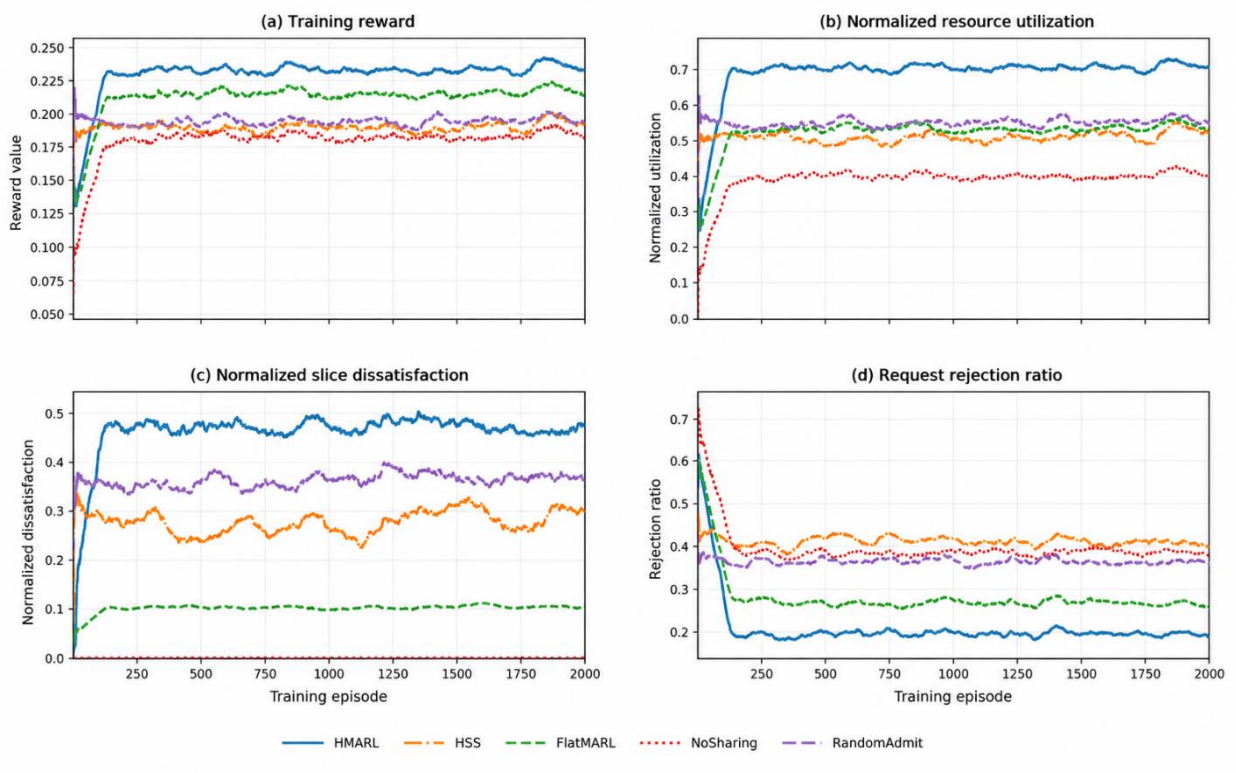


Fig. 1 Training convergence of the evaluated online methods over 2000 training episodes.

Taken together, the four metrics indicate that HMARL reaches a stable operating regime after the initial training phase. The observed performance does not result from minimizing slice dissatisfaction alone; rather, HMARL accepts a higher sharing burden while achieving substantially higher resource utilization and a lower request rejection ratio. This trade-off is reflected in the higher and stable training reward observed during the later stages of training.

6-2- Comparative Performance Across Traffic Scenarios

Fig. 2 compares HMARL with the considered learning-based, heuristic, random, non-sharing, and offline reference methods. The reported values are aggregated across the evaluated traffic scenarios and include the objective value, normalized resource utilization, normalized slice dissatisfaction, and request rejection ratio.

As shown in Fig. 2(a), HRASS achieves the highest objective value at approximately 0.25, followed closely by HMARL at about 0.24 and the w/o Meta variant at about 0.23. The small difference between HMARL and HRASS is notable because HRASS operates offline with broader knowledge of the requests over the scheduling horizon, whereas HMARL makes its decisions online using only currently available information.

Fig. 2(b) shows a similar trend in resource utilization. HRASS achieves the highest normalized utilization, approximately 0.76, followed by HMARL at about 0.73 and Greedy at about 0.68. The lower utilization of the w/o Meta variant highlights the contribution of adaptive meta-level borrowing coordination. As shown in Fig. 2(c), the higher utilization of HMARL is accompanied by a higher normalized slice dissatisfaction of approximately 0.51, reflecting its more active use of inter-slice borrowing. In contrast, No-Sharing maintains near-zero dissatisfaction because borrowing is disabled, but this is accompanied by substantially lower utilization and a higher rejection ratio.

Figure 2(d) further shows that HRASS achieves the lowest rejection ratio, approximately 0.32, followed by Greedy at about 0.35 and HMARL at about 0.37. The w/o Meta variant reaches approximately 0.44, while the remaining methods exhibit higher rejection levels.

Overall, HMARL provides a favorable trade-off among resource utilization, slice dissatisfaction, and request rejection, while achieving an objective value close to the offline HRASS reference. These results also show that adaptive meta-level coordination improves resource utilization and request admission at the cost of a controlled increase in inter-slice sharing.

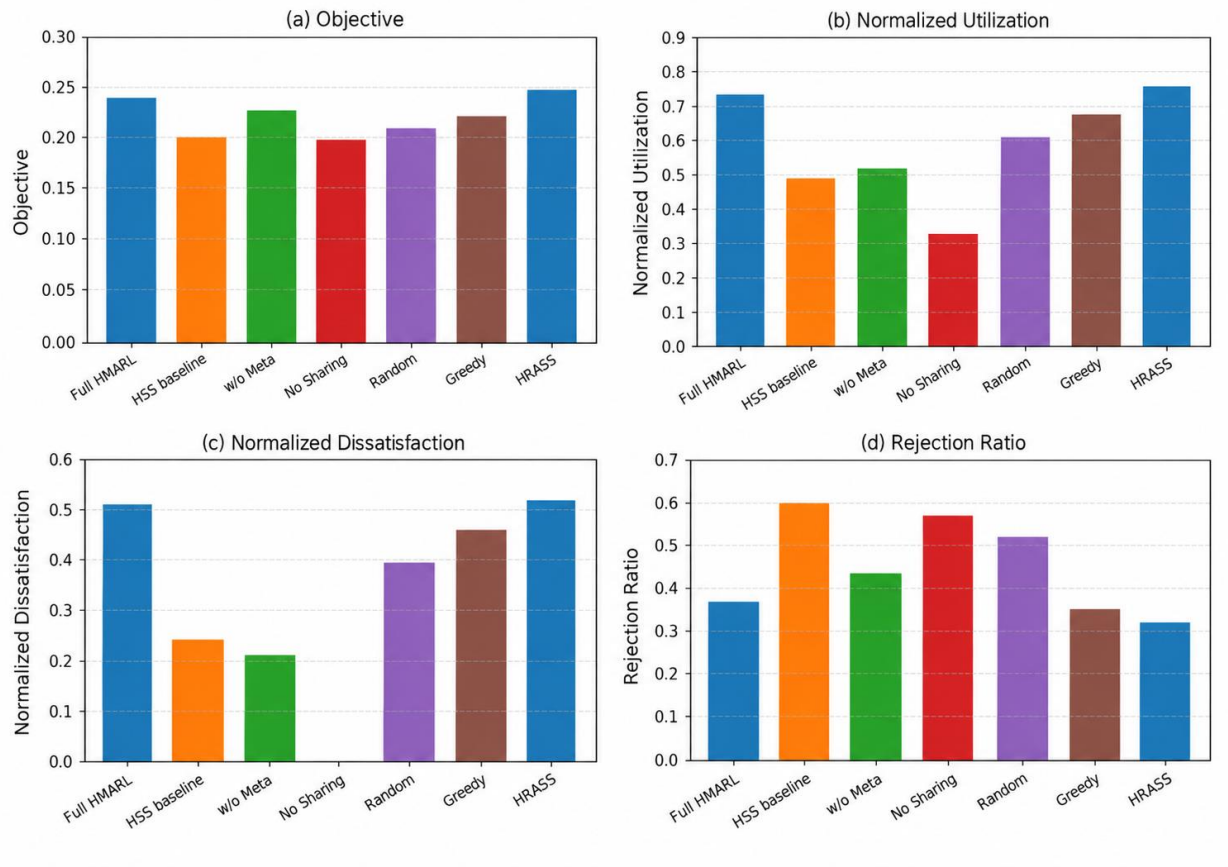


Fig. 2 Comparison of the evaluated methods across the traffic scenarios.

7- Conclusion

In this paper, we proposed a Hierarchical Multi-Agent Reinforcement Learning (HMARL) framework for online soft and safe network slicing. The framework combines slice-level admission decisions with a higher-level meta-controller that adaptively regulates inter-slice borrowing, while a deterministic safety layer ensures that only feasible resource allocations are executed. The experimental results show that HMARL converges to a stable policy and achieves a favorable balance among resource utilization, slice dissatisfaction, and request rejection. Compared with the considered learning-based, heuristic, random, and non-sharing alternatives, HMARL provides high resource utilization and competitive objective values while maintaining a relatively low rejection ratio. The comparison with the w/o Meta variant further demonstrates the benefit of meta-level borrowing coordination.

Future work will investigate more heterogeneous and time-varying traffic conditions, richer slice-specific service requirements, and adaptive multi-objective resource-management strategies. Further extensions may also consider larger and more dynamic network environments and mechanisms for improving the interpretability of hierarchical decisions in practical network slicing deployments.

References

[1] Chowdhury, M.Z., et al., 6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions. IEEE Open Journal of the Communications Society, 2020. 1: p. 957-975.

[2] Afolabi, I., et al., Network slicing and softwarization: A survey on principles, enabling technologies, and solutions. IEEE Communications Surveys & Tutorials, 2018. 20(3): p. 2429-2453.

[3] Vassilaras, S., et al., The algorithmic aspects of network slicing. IEEE Communications Magazine, 2017. 55(8): p. 112-119.

[4] Khan, L.U., et al., Network slicing: Recent advances, taxonomy, requirements, and open research challenges. IEEE Access, 2020. 8: p. 36009-36028.

[5] Jiang, W., et al., The road towards 6G: A comprehensive survey. IEEE Open Journal of the Communications Society, 2021. 2: p. 334-366.

[6] Sasan, Z., et al. Joint network slicing, routing, and in-network computing for energy-efficient 6g. in 2024 IEEE Wireless Communications and Networking Conference (WCNC). 2024. IEEE.

[7] Gonzalez, A.J., et al. The isolation concept in the 5G network slicing. in 2020 European Conference on Networks and Communications (EuCNC). 2020. IEEE.
[8] Sasan, Z. and S. Khorsandi. Slice-aware resource calendaring in cloud-based radio access networks. in 2022 30th International Conference on Electrical Engineering (ICEE). 2022. IEEE.
[9] Zhang, S., et al., Hierarchical soft slicing to meet multi-dimensional QoS demand in cache-enabled vehicular networks. IEEE Transactions on Wireless Communications, 2020. 19(3): p. 2150-2162.
[10] Wang, T., et al., LinkSlice: Fine-grained network slice enforcement based on deep reinforcement learning. IEEE Journal on Selected Areas in Communications, 2022. 40(8): p. 2378-2394.
[11] Zhang, H., et al. A hard and soft hybrid slicing framework for service level agreement guarantee via deep reinforcement learning. in 2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring). 2022. IEEE.
[12] Kim, Y. and H. Lim, Multi-agent reinforcement learning-based resource management for end-to-end network slicing. IEEe Access, 2021. 9: p. 56178-56190.
[13] Sulaiman, M., et al. Multi-agent deep reinforcement learning for slicing and admission control in 5G C-RAN. in NOMS 2022-2022 IEEE/IFIP network operations and management symposium. 2022. IEEE.
[14] Mason, F., G. Nencioni, and A. Zanella. A multi-agent reinforcement learning architecture for network slicing orchestration. in 2021 19th Mediterranean Communication and Computer Networking Conference (MedComNet). 2021. IEEE.
[15] Sasan, Z. and S. Khorsandi, Balancing resource utilization and slice dissatisfaction through dynamic soft slicing for 6G wireless networks. Scientific Reports, 2025. 15(1): p. 22987.
[16] Li, J., et al., A hierarchical soft RAN slicing framework for differentiated service provisioning. IEEE Wireless Communications, 2020. 27(6): p. 90-97.
[17] Shi, W., et al., Two-level soft RAN slicing for customized services in 5G-and-beyond wireless communications. IEEE Transactions on Industrial Informatics, 2021. 18(6): p. 4169-4179.
[18] Ding, J., M. Zheng, and H. Yu, Soft Resource Slicing for Industrial Mixed Traffic in 5G Networks. IEEE/CAA Journal of Automatica Sinica, 2025. 12(2): p. 463-465.
[19] Kaytaz, U., F. Sivrikaya, and S. Albayrak. Hierarchical deep reinforcement learning based dynamic RAN slicing for 5G V2X. in 2021 IEEE Global Communications Conference (GLOBECOM). 2021. IEEE.
[20] Shao, Y., et al., Graph attention network-based multi-agent reinforcement learning for slicing resource management in dense cellular network. IEEE Transactions on vehicular Technology, 2021. 70(10): p. 10792-10803.
[21] Toumi, N., M. Bagaa, and A. Ksentini. Hierarchical multi-agent deep reinforcement learning for SFC placement on multiple domains. in 2021 IEEE 46th Conference on Local Computer Networks (LCN). 2021. IEEE.
[22] Chen, X., et al. Hierarchical Meta-Reinforcement Learning for Resource-Efficient Slicing in O-RAN. in GLOBECOM 2023-2023 IEEE Global Communications Conference. 2023. IEEE.
[23] Akyıldız, H.A., et al., Hierarchical reinforcement learning based resource allocation for RAN slicing. IEEE Access, 2024. 12: p. 75818-75831.
[24] Phyu, H.P., D. Naboulsi, and R. Stanica, ICE-CREAM: multI-agent fully CooperativE deCentRalizEd frAMework for Energy Efficiency in RAN Slicing. IEEE Transactions on Network and Service Management, 2025.
[25] Hu, T., et al. Inter-cell slicing resource partitioning via coordinated multi-agent deep reinforcement learning. in ICC 2022-IEEE International Conference on Communications. 2022. IEEE.
[26] Zhou, G., et al., Joint multi-objective optimization for radio access network slicing using multi-agent deep reinforcement learning. IEEE Transactions on Vehicular Technology, 2023. 72(9): p. 11828-11843.
[27] Zhou, H., M. Elsayed, and M. Erol-Kantarci. RAN resource slicing in 5G using multi-agent correlated Q-learning. in 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). 2021. IEEE.

Journal of Information Systems and Telecommunication
Vol.14, No.4, July-September 2026, 238-244
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Assessing the Usability of a Comprehensive Research Management System: A Mixed-Use Study in Iran

Mehdi Beyrami Jam[1], Mohammad Sadegh Sohrabi[2*], Mohammad Babamiri[3], Arezoo Farhadi[4]

[1].Department of Health in Emergency and Disaster, School of Public Health, University of Medical Sciences, Hamadan, Iran.
[2].Occupational Health and Safety Research Center, Institute of Health Sciences and Technology, Hamadan University of Medical Sciences, Hamadan, Iran.
[3].Department of Ergonomics, School of Public Health, Research Center for Health Sciences, Institute of Health Sciences and Technology, Hamadan University of Medical Sciences, Hamadan, Iran.
[4].Department of Medical Library and Information Science, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran.

Received: 15 Feb 2026/ Revised: 04 Aug 2026/ Accepted: 09 Sep 2026

Abstract

Usability is a key dimension of the quality of research information systems. This study aimed to examine the usability of the "Pazhuheshyar" research affairs management system and explore users' experiences and challenges when interacting with it across several medical universities in Iran. A convergent parallel mixed-methods design was employed. In the quantitative phase, 169 users (76 men and 93 women) completed a questionnaire evaluating the usability of web-based software, and independent t-tests were used to examine differences by gender and user role. In the qualitative phase, 16 semi-structured interviews were conducted using purposive sampling and analyzed through content analysis. The mean Relative Usability score was 33.19%, indicating a low to moderate level of usability. The lowest subscale score was related to Ease of Use (23.82%), while the highest score was observed for Content/Output Quality (42.40%). Women reported higher usability scores than men across most domains, although no significant differences were found for System Terminology and Information and System Capabilities. Similarly, officers and managers reported higher scores than researchers on most scales, with no significant differences in System Terminology and Information, Personalization, and Error Management. Qualitative analysis identified five main categories: complex navigation pathways and similar menu options; system errors and data storage problems; insufficient training and user support; strengths related to documentation and national system integration; and suggested improvements such as simplifying processes, redesigning search and reporting functions, providing personalized dashboards, and developing a mobile version. Despite strengths in documentation and integration, limitations in ease of use and learnability appear to hinder optimal efficiency and user satisfaction. Implementing user-centered workflow redesign, improving error handling and feedback mechanisms, providing structured training, and continuously monitoring user experience may enhance system adoption, user satisfaction, and productivity.

Keywords: Ergonomics; Research Management System; User-Computer Interface; Usability.

1- Introduction

Usability is widely recognized as a fundamental dimension of information system quality. According to ISO 9241-11, usability refers to the extent to which specified users can achieve defined goals within a particular context of use effectively, efficiently, and with satisfaction [1]. In university research management systems—designed to support the entire research lifecycle from proposal submission and review to funding, implementation, and project closure—usability plays a critical role in shaping researchers' productivity, system acceptance, and overall user satisfaction. From the perspective of human–computer interaction and cognitive ergonomics, systems designed according to user-centered principles can significantly enhance learnability, interaction efficiency, and perceived usefulness [2,3].

Existing research on research management and academic information systems has commonly relied on structured usability evaluation frameworks and mixed-methods approaches. Several studies combining heuristic evaluation based on Nielsen's ten heuristics with the System Usability Scale (SUS) have identified recurring usability issues

✉ **Mohammad Sadegh Sohrabi**
ms.sohrabi@umsha.ac.ir

related to interface clarity, feedback mechanisms, color usage, and command design. Although these studies often reported acceptable levels of user competence and general satisfaction, they consistently recommended improvements in visual clarity, feedback provision, and interface guidance [4]. In contrast, evaluations of other university systems have revealed substantially lower usability levels, with SUS scores as low as 47 accompanied by high levels of user dissatisfaction and perceived complexity [5]. Similarly, even systems reporting relatively high usability scores (e.g., 81.25) have demonstrated limitations in search functionality and information presentation, indicating the need for improved search visualization and clearer evaluation criteria [6].

Beyond system-level usability scores, previous research has also highlighted the broader relationship between usability and user outcomes. For instance, staff surveys have shown a positive association between perceived system usability and satisfaction with online communication environments [7,8]. Expert heuristic evaluations have further revealed numerous usability problems, particularly those related to the principle of "recognition rather than recall," which directly affects users' cognitive workload and task performance [8,9]. Likewise, student-based evaluations have documented large numbers of usability issues, most commonly associated with aesthetic and minimalist design principles as well as limitations in user control and navigation freedom [10]. Methodological research has also emphasized the effectiveness of hybrid usability evaluation approaches in higher-education management systems, with users generally reporting acceptance of these combined assessment frameworks [11,12].

Evidence from the Iranian context reflects similar patterns. Bahador et al., in an exploratory evaluation of a research management system, reported that although some fundamental functional requirements were addressed, major usability limitations remained, particularly in relation to user compatibility, interface simplicity, system guidance, and feedback mechanisms, highlighting the need for stronger user-centered design practices [10]. Likewise, usability evaluations of the Ministry of Health Examination System identified moderate usability scores together with notable challenges in navigation structure, access to key functions, and error feedback mechanisms [12,13].

Collectively, these findings emphasize that the success of research management systems depends not only on functional capabilities but also on the quality of user interaction, system learnability, and alignment with human-centered design principles.

Another study reported an inverse relationship between system usability and users' mental workload, measured using the NASA-TLX scale. In that evaluation, complex navigation structures and insufficient feedback mechanisms were associated with increased cognitive load, highlighting the importance of user-centered redesign in academic information systems [13,14]. Consistent evidence from both international and national studies indicates that recurring usability problems—such as complex navigation pathways, high memory demands, inadequate feedback, inconsistent terminology, and non-minimalist interface design—can reduce usability scores, increase mental workload, and ultimately weaken system acceptance among users [13–17]. Conversely, continuous monitoring of user experience combined with iterative design improvements based on quantitative and qualitative feedback has been associated with improved usability outcomes in academic information systems [16,17].

Despite the expanding body of research on usability evaluation in higher-education information systems, several important gaps remain. Many previous investigations have relied primarily on quantitative usability scales or expert-based heuristic evaluations, whereas fewer studies have adopted integrated approaches that combine standardized usability measurement with in-depth exploration of users' experiences. Furthermore, empirical evidence regarding the usability of research management systems in medical universities remains limited, particularly within the Iranian context. Addressing these gaps requires systematic evaluations that simultaneously assess usability performance and explore user experiences to generate actionable design improvements.

Hamadan University of Medical Sciences currently uses the "Pazhuheshyar" platform to manage research activities. Preliminary observations suggest that a systematic evaluation grounded in local user experiences could help identify both system strengths and usability limitations, clarify stakeholders' needs, and inform practical improvements. To date, many existing studies in similar contexts have been limited to purely quantitative assessments or exploratory analyses, and relatively few investigations within medical universities have employed a convergent mixed-methods design that simultaneously quantifies usability and examines users' lived experiences.

Accordingly, the present study evaluates the usability of the Pazhuheshyar research management system at Hamadan University of Medical Sciences using a convergent mixed-methods approach. By integrating quantitative usability measurement with qualitative exploration of user experiences, the study seeks to provide a comprehensive understanding of system usability and to identify practical opportunities for improving user interaction and system performance.

The study addresses the following research questions:

1) What is the overall usability level of the Pazhuheshyar research management system?
2) Are there differences in perceived usability across user roles and gender?
3) What usability challenges and barriers do users encounter while interacting with the system?

Through the integration of quantitative and qualitative findings, this study aims to provide a deeper understanding of the system's usability strengths and limitations and to offer evidence-based recommendations for improving user experience and system efficiency.

2- Materials and Methods

This study employed a cross-sectional convergent mixed-methods design, in which the quantitative and qualitative components were conducted in parallel and subsequently integrated during the discussion and interpretation phase. The target population consisted of users of the university's Pazhuheshyar research management system, including faculty members and system officers (administrators/support staff), who had registered or followed at least one research activity or project in the system within the previous 12 months. Participants were eligible for inclusion if they (1) had an active user account in the system, (2) had experience using the system during the past 12 months, and (3) provided informed and voluntary consent to participate. Exclusion criteria were (1) incomplete completion of the questionnaire in the quantitative phase and (2) withdrawal during the qualitative interviews.

In both study components, a stratified sampling approach was used based on user roles (faculty/researcher, research officer, or system manager). Based on a prior estimation [15] with a standard deviation of 15.55 for the usability index and a target population of approximately 521 users, the required sample size for the quantitative phase was calculated to be 162 participants (415 faculties/researchers and 106 officers/managers). To account for potential non-response, the questionnaire was distributed to the entire eligible population (n = 521). For the qualitative phase, purposeful sampling was applied with maximum variation in user roles, faculties, and levels of digital literacy, and the process continued until data saturation was reached.

2-1- Instrument

To evaluate usability, the Web-Based Application Usability Evaluation Questionnaire was used. This instrument includes 50 Likert-scale items covering the main dimensions of effectiveness, efficiency, safety/error management, features/capabilities, and learnability/memorability. The total and subscale scores were calculated and normalized on a 0–100 scale, where higher scores indicate better usability [15]. The validity and reliability of this instrument have been confirmed in previous similar studies [15, 18]. For data collection in the quantitative phase, the original questionnaire text was directly converted into an online format, which was distributed via SMS Center services to the mobile phones of all target population members for completion. The internal consistency of the usability questionnaire was evaluated using Cronbach's alpha. The overall reliability coefficient obtained for the present sample was 0.82, indicating satisfactory internal consistency of the instrument.

In the qualitative phase, semi-structured interviews were conducted with 16 participants, whose characteristics are presented in Table 1. The interviews explored three main areas: (1) user experience and satisfaction, (2) problems and barriers, and (3) suggestions for improvement. Data collection continued until theoretical/data saturation was achieved. With the participants' consent, interviews were audio-recorded and transcribed verbatim. The data were analyzed using MAXQDA qualitative analysis software, through which initial codes and categories were extracted. To enhance the trustworthiness of the qualitative findings, several strategies were employed. Credibility was supported through careful transcription and repeated review of interview data. Coding and categorization were conducted independently by two researchers, and discrepancies were resolved through discussion until consensus was achieved. Data saturation was considered reached when no new themes emerged from subsequent interviews. These procedures helped strengthen the credibility and dependability of the qualitative analysis.

The integration of quantitative and qualitative findings provided complementary insights into the usability of the Pazhuheshyar system. While quantitative results identified the dimensions with the lowest usability scores, the qualitative data helped explain the underlying reasons behind these findings. This triangulation of evidence improved the interpretability of results and allowed a deeper understanding of users' interaction difficulties with the system.

2-2- Data Analysis

Descriptive statistics (mean, standard deviation, frequency, and percentage) and inferential analyses (independent t-test, $\alpha = 0.05$) were used to analyze the quantitative data. Reliability of the usability questionnaire was assessed using Cronbach's alpha, indicating acceptable internal consistency for the overall scale and its subscales. Prior to conducting parametric tests, assumptions of normality and homogeneity of variance were examined, and effect size measures were calculated to support a more meaningful interpretation of group differences. Qualitative data were analyzed using conventional content analysis based on the approach proposed by Graneheim and Lundman. The analysis involved identifying meaning units, generating initial codes, organizing subcategories, and developing broader categories and themes. Coding was performed independently by two researchers, and disagreements were resolved through discussion and consensus [19].

3- Results

3-1- Quantitative Findings

The participants in this study included 169 people, 76 (45%) were male and 93 (55%) were female in the quantitative section and 16 (6 males and 10 females) in the qualitative section. The demographic characteristics of the participants are shown in Table 1.

Table 1. Demographic Characteristics of Participants (Quantitative and Qualitative)

	Group/ Level	Quantitative (n = 169)	%	Qualitative (n = 16)	%
Gender	Male	76	45	6	37.5
	Female	93	55	10	62.5
Age	—	41.24 ± 8.81	—	40.22 ± 6.50	—
Position	Researcher	141	83	11	68.8
	administrative user	28	17	5	31.2

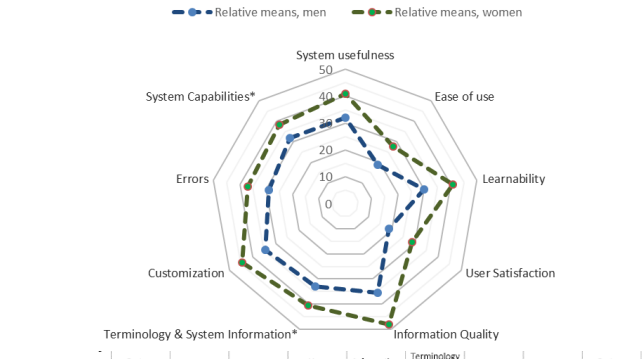
The preliminary results regarding the relative usability score are presented in Table 2. Considering the relative nature of each subscale, it was observed that most mean scores were approximately 34%, indicating a substantial deficiency in the level of usability reported by participants. The lowest mean score was related to ease of use (mean = 23.82), whereas the highest mean score (mean = 42.40) corresponded to content/output display on the screen.

Table. 2. Relative (0-100) usability scores and its subscales (n=169)

	Mean	SD
System usefulness	37.01	28.07
Ease of use	23.82	22.37
Learnability	35.97	24.89
User Satisfaction	24.47	24.93
Information (Content/Output) Quality	42.40	26.40
Terminology and System Information	37.20	25.61
Customization	39.94	26.28
Errors	33.28	23.96
System Capabilities	35.50	23.71
Relative Usability (overall)	33.19	22.59

Higher scores reflect better usability (0–100). Subscale and overall values are reported as corrected relative scores.

Figure 1 presents the differences in mean Relative Usability subscale scores between men and women. As shown, all subscales—except Terminology & System Information and System Capabilities—exhibit statistically significant differences, with women scoring higher on average than men.



	System usefulness	Ease of use	Learnability	User Satisfaction	Information Quality	Terminology & System Information*	Customization	Errors	System Capabilities*
Relative means, men	32.07	18.88	30.02	19.01	35.46	33.06	34.54	28.95	32.00
Relative means, women	41.04	27.85	40.83	28.92	48.06	40.59	44.35	36.83	38.36

Figure 1. Mean difference in relative usability and subscales between men and women. * Not significant difference (P>0.05)



	System usefulness	Ease of use	Learnability	User Satisfaction	Information Quality	Terminology & System Information*	Customization*	Errors*	System Capabilities
Relative means, reserchers	34.78	21.10	32.58	21.70	39.82	35.51	38.21	32.21	33.46
Relative means, admins	48.21	37.50	53.01	38.39	55.36	45.76	48.66	38.69	45.79

Figure 2. Mean difference in relative usability and subscales between researchers and admins. * Not significant difference (P>0.05)

Figure 2 shows the differences in mean scores across the nine usability subscales between researchers and system officers/managers. The results indicate statistically significant differences for all subscales except Terminology & System Information, Customization, and Errors. Across all subscales, the relative mean scores of officers/managers were higher than those of researchers.

3-2- Qualitative Findings

Based on sixteen interviews with system officers (administrators/support staff) and user researcher of the Pazhuheshyar system, Figure 3 summarizes five categories: (1) Interaction complexity; (2) technical/system issues; (3) support and training; (4) system strengths; and (5) Usability Improvements. Across interviews, usability difficulties were pervasive: cluttered menus, long and complex pathways for project registration, and highly similar options that led to user errors. For example, in the topic of Interaction complexity, interviewee number 5 stated, "Notification system: SMS, emails, and notifications that

are sometimes incomplete or unclickable". Confusion in the review workflow and lack of coordination across system modules were also repeatedly noted. Technically, participants reported login errors, unreliable saving of information, the need to clear the browser cache, and mismatches between research centers' processes and the system's forms—issues that consumed time and eroded trust. Reporting and search functions were frequently highlighted as especially problematic.

Regarding support and training, initial instruction was widely viewed as insufficient; many users learned the system through trial and error or frequent contact with support staff. While workshops and a ticketing system were said to alleviate some challenges, others emphasized weak support and reliance on individual experience. Users nonetheless identified strengths, notably documentation and archiving of research records (proposals, theses), notification/SMS features, and linkage to ministry-level macro systems. Suggested improvements converged on streamlining workflows and removing nonessential options, adding dashboard personalization, redesigning search, strengthening notifications/alerts, creating an integrated researcher database, and developing a dedicated version. Overall, while documentation and national integration are valued, persistent usability, technical, and training gaps indicate a need for user-centered redesign.

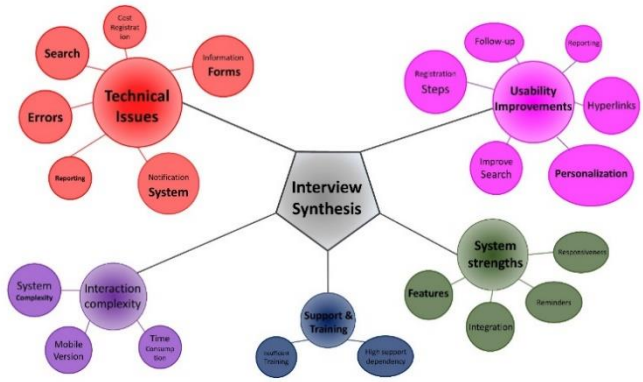


Figure 3. Challenges and solutions for improving system usability

3-3- Integration and Comparison of Quantitative and Qualitative Findings

A synthesis of the quantitative and qualitative results reveals a coherent pattern across both components. Both lines of evidence highlight major weaknesses in Ease of Use, Learnability, and Error Management, indicating that users' practical challenges—such as complex navigation and recurrent system errors—directly mirror the lowest-scoring dimensions identified in the quantitative data. At the same time, convergence was observed in the system's strengths: both datasets underscored positive perceptions of Content and Output Quality, particularly regarding documentation, data integration, and notification features. This consistency suggests that while the system provides valuable informational functions, its interface design and interaction flow remain key obstacles to efficient use. Group-level patterns were also aligned, with officers and managers—who had greater exposure and training—reporting higher usability than researchers, who expressed more difficulty and dependence on trial-and-error use.

In summary, the merged interpretation indicates that the system's usability limitations are primarily rooted in design and interaction issues rather than information content. Implementing user-centered redesign, enhanced feedback mechanisms, and structured training is therefore essential to improving overall usability and user satisfaction.

4- Conclusions

From a usability perspective, the low scores observed for ease of use and learnability suggest challenges related to interface complexity, workflow structure, and elevated cognitive load. Complex navigation paths, layered menus, and limited system feedback likely increase users' mental workload and reduce interaction efficiency—an interpretation consistent with human–computer interaction principles linking unclear navigation and insufficient feedback to lower perceived usability and higher cognitive demand.

In the quantitative phase, 169 users (76 men and 93 women) evaluated the Pazhuheshyar system on a 0–100 scale, yielding an overall mean of approximately 33, indicative of low-to-moderate usability. Ease of Use received the lowest score (23.82%), whereas Content/Output Quality scored highest (42.4%). Women generally reported higher usability scores than men, and officers/managers outperformed researchers across most dimensions. The qualitative analysis of 16 interviews further contextualized these findings, revealing five main categories: (1) complex navigation structures; (2) system and storage errors; (3) inadequate training and support; (4) strengths in documentation and data integration; and (5) suggested improvements, including workflow simplification, enhanced search functionality, and greater interface personalization.

Weaknesses in ease of use and learnability alongside comparatively better information output align with prior reports and a closely related study of research systems; many evaluations find stronger performance in documentation and data integration, but overall usability suffers due to complex navigation, inconsistent terminology, and high memory load [6]. Differences favoring system

administrators over researchers have been attributed in the literature to greater procedural familiarity, access to specific pathways, and more targeted training—an interpretation consistent with the present findings [11, 13]. Moreover, the documented link between lower usability and greater mental workload/time loss corresponds with our qualitative themes of navigational confusion and recurrent errors [4, 9]. Our results also accord with Kazemi et al. [14] on the research management system at Shiraz University of Medical Sciences, where usability scores for submission, review, and monitoring were below 50, with negative correlations between usability and "mental effort" and "errors"—i.e., as usability decreases, mental effort and errors increase.

The convergence of quantitative and qualitative evidence yields a coherent picture: low Ease of Use and Learnability in the metrics mirrors qualitative reports of long pathways, crowded menus, similar options, and reliance on trial-and-error to learn the system. Conversely, the relatively better Content/Output Quality aligns with reported strengths in documentation, archiving, notifications, and linkage to national platforms. Role differences favoring officers/managers can largely be explained by more targeted training/support and more frequent use patterns, while researchers' complaints about difficult search/reporting clarify their lower scores. Finally, gender differences favoring women (at least descriptively) may reflect usage patterns or sample composition and warrant controlled analyses in future research [5, 16, 20].

To improve Pazhuheshyar usability, we propose an integrated strategy starting with task-oriented redesign: reduce process steps, remove low-utility options, and organize workflows around frequent scenarios (registration, revision, review) to lower cognitive load. In parallel, strengthen error management and feedback so messages are informative/actionable and undo/cancel is available. Enhance information retrieval by redesigning the search page, adding filters, and offering predefined report templates [21, 22]. To accommodate diverse needs, provide useful, customizable dashboards for researchers and officers/managers, supported by structured training and support (in-system micro learning modules, step-by-step guides, and an integrated ticketing system). Ongoing UX (user experience) monitoring via remote measurement and behavioral indicators (task time, completion rate, error rate) with iterative improvement cycles should sustain quality; attention to accessibility and a mobile version for light tasks/follow-ups can broaden effective use. Key limitations include the cross-sectional design, reliance on self-report in the quantitative tool, and absence of behavioral measures alongside the questionnaire, which affect generalizability. Future phases should incorporate task-based usability testing, process data, and multivariable analyses to control for confounders.

This study indicates that although the Pazhuheshyar system provides important strengths in documentation and data integration, it faces persistent challenges in user experience, technical performance, and user training. Low scores in ease of use and learnability, recurring system and saving errors, and inconsistencies across forms and workflows appear to reduce efficiency and contribute to user dissatisfaction. The lack of structured training further exacerbates these issues, leaving many users dependent on trial-and-error or informal support. Addressing these challenges requires a user-centered redesign of the system. Key priorities include aligning the interaction architecture with user needs, simplifying workflows, improving search and reporting functions, and resolving technical errors. In parallel, clear guidelines, structured training programs, and an integrated support system are needed to strengthen user competence and system acceptance. Additional improvements—such as dashboard personalization and a mobile interface for lightweight tasks—could further accommodate diverse user roles and working contexts. Overall, the effectiveness of research management systems depends not only on technical functionality but also on their alignment with users' needs and interaction patterns. Accordingly, the proposed improvement roadmap emphasizes user-centered redesign, continuous monitoring of user experience, and phased implementation of usability enhancements to support sustainable improvements at institutional and national levels.

Competitive Interests

The authors declare no conflicts of interest related to the subject matter of this research or its findings.

Ethics Approval

The study protocol was approved by the Ethics Committee of Hamadan University of Medical Sciences (IR.UMSHA.REC.1404.348). Informed consent, confidentiality, and the right to withdraw at any time were observed. Data were reported in aggregate and without identifying information.

Acknowledgments

The authors would like to thank the entire Vice President for Research and Technology and its staff. This project was also conducted in 2025 with financial support from the Research and Technology division of Hamadan University of Medical Sciences (Project number 140405144056).

References

[1] ISO 9241-11:2018. Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts, International Organization for Standardization, 2018.

[2] P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, Usability Evaluation in Industry. Taylor & Francis, 1996.

[3] M. M. Yusof, J. Kuljis, A. Papazafeiropoulou, and L. K. Stergioulas, "An evaluation framework for Health Information Systems: human, organization and technology-fit factors (HOT-fit)," (in eng), Int J Med Inform, vol. 77, no. 6, pp. 386-98, Jun 2008.

[4] P. J, I. Pendidikan, Lisnawita, and C. Loneli, "Exploring Research and Service Information System Usability by Heuristic Evaluation as a Compelement of System Usability Scale," presented at the Jurnal Penelitian Pendidikan IPA, 2023.

[5] P. Edilson, M. Érica, V. José, M. Cristiano, and T. Daniela, "Avaliação da Usabilidade do Sistema de Informação e Gestão de Projetos (SIGProj) e da Aceitação de uma Proposta na Extensão Universitária," Anais do Computer on the Beach, vol. 14, pp. 272-279, 2023.

[6] S. C et al., "Empowering Researchers to Query Medical Data and Biospecimens by Ensuring Appropriate Usability of a Feasibility Tool: Evaluation Study," JMIR Human Factors, vol. 19, no. 10, p. e43782, 2022.

[7] A. M. S. Mohd, S. Ali, N. S. A. Mohd, and H. Hasrul, "The Importance of Usability Features in Enhancing Online Communication Satisfaction," Malaysian Journal of Communication, vol. 32, no. 1, pp. 1-15, 2016.

[8] K. Qureshi, Alhajri, Hessah , El-barr, Mostafa Abd, "Design and Analysis of a Secure Intelligent Blood Management Information System," Journal of Information Systems and Telecommunication (JIST), vol. 12, no. 48, pp. 300-312, 2025.

[9] S. Etemadinezhad, E. Amouzadeh, and J. Charati, "Investigating the Usability of the Integrated Research Automation System (SYAT): A Heuristic Evaluation," Iranian Journal of Ergonomics, vol. 10, no. 1, pp. 46-55, 2023.

[10] F. Bahador, F. Salehi, E. Esmaeeli, A. Akbari, and A. Sabahi, "Evaluation the Usability of Hamava System Using Heuristic Evaluation," Frontiers in Health Informatics, vol. 9, no. 12, pp. 11045–11052, 2023.

[11] G.-P. Patricio, F.-P. Félix, and U.-U. Pilar, "Hybrid Procedure for Measuring Usability of Management Systems in Higher Education," Iberian Conference on Information Systems and Technologies, 2019.

[12] M. D. E. Fernanda, E. Leandro, and P. B. Denise, "Usability of the “Systematic Review Support” computer system: a methodological study," Revista Latino-Americana de Enfermagem, 2024.

[13] J. Nikokaran, S. A. Fatemi Aghda, H. Dehghan, M. Mirfakhraei, M. M. Poornematy, and M. Mirfakhraei, "Usability Evaluation of the Ministry of Health Exam System: Viewpoint the Users," (in eng), Journal of Health and Biomedical Informatics, Original Article vol. 11, no. 3, pp. 176-185, 2024.

[14] R. Kazemi, Z. Rasouli kahaki, and A. S. Jafari Roodbandi, "The relationship between the usability of research centers system and mental workload caused by its interaction with users of shiraz university of medical sciences," (in eng), Iran Occupational Health Journal, Research vol. 18, no. 1, pp. 274-284, 2021.

[15] J. Faradmal and J. Keshvari Kamran, "The validity and reliability of an usability assessment tool for a web-based software," (in eng), Iranian Journal of Ergonomics, Research vol. 2, no. 3, pp. 57-69, 2014.

[16] S. Etemadinezhad, E. Amouzadeh, and J. Yazdani Charati, "Investigating the Usability of the Integrated Research Automation System (SYAT): A Heuristic Evaluation," Iranian Journal of Ergonomics, vol. 10, no. 1, pp. 46-55, 2022.

[17] A. Bangor, P. T. Kortum, and J. T. Miller, "An Empirical Evaluation of the System Usability Scale," International Journal of Human–Computer Interaction, vol. 24, no. 6, pp. 574-594, 2008.

[18] I. Dianat, Z. Ghanbari, and M. AsghariJafarabadi, "Psychometric properties of the persian language version of the system usability scale,", Health Promot Perspect, vol. 4, no. 1, pp. 82-9, 2014.

[19] U. H. Graneheim and B. Lundman, "Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness," (in eng), Nurse Educ Today, vol. 24, no. 2, pp. 105-12, Feb 2004.

[20] D. dmp, Mahjur, Ali, "Concatenating Approach: Improving the Performance of Data Structure Implementation," Journal of Information Systems and Telecommunication (JIST), vol. 6, no. 21, pp. 18-24, 2018.

[21] M. Beyramijam, M. S. Sohrabi, M. Babamiri, and A. Farhadi, "Usability of Research and Educational Systems: A Policy Brief,", Iranian Journal of Ergonomics, vol. 13, no. 4, 2026.

[22] H. hamidi, "Selecting Enterprise Resource Planning System Using Fuzzy Analytic Hierarchy Process Approach," Journal of Information Systems and Telecommunication (JIST), vol. 3, no. 12, pp. 1-10, 2015.

Journal of Information Systems and Telecommunication
Vol.14, No.4, July-September 2026, 245-267
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

A Hybrid Entity-Centric Approach to Persian Pronoun Resolution with BERT Enhancement

Hassan Haji Mohammadi[1], Alireza Talebpour[2*], Ahmad Mahmoudi-Aznaveh[3], Samaneh Yazdani[1]

[1].Department of Computer Engineering, NT.C., Islamic Azad University, Tehran, Iran
[2].Department of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
[3].Cyberspace Research Institute, Shahid Beheshti University, Tehran, Iran

Received: 31 Jan 2025/ Revised: 04 Sep 2026/ Accepted: 25 Oct 2026

Abstract

Pronoun resolution, a critical aspect of coreference resolution in natural language processing, is essential for accurate text understanding and generation. This paper introduces a hybrid Persian pronoun resolution model that leverages rule-based and machine-learning approaches. Seven high-precision rule-based modules, tailored explicitly for Persian linguistic nuances, are combined with a random forest classifier enhanced with BERT contextual embeddings and cluster-level features. This approach effectively addresses the challenges of pronoun resolution, surpassing the performance of existing Persian systems on two corpora: the newly developed Mehr corpus (400 documents) and the established RCDAT corpus. Integrating BERT embeddings and cluster-level features significantly contributes to the model's superior performance, achieving a 4.68 F1-score improvement on Mehr and a 4.01 F1-score improvement on RCDAT. The proposed hybrid model demonstrates a practical and interpretable solution, contributing to Persian natural language processing research and applications.

Keywords: Coreference Resolution; Natural Language Processing; Pronoun Resolution; Corpus Development; BERT; Transformer Models.

1- Introduction

Coreference resolution, a fundamental task within natural language processing (NLP), identifies and groups mentions within a text that refers to the same real-world entity [1, 2]. Accurate coreference resolution is essential for various NLP applications, including information extraction, text summarization, machine translation, question answering, and sentiment analysis [3-7]. The coreference resolution task encompasses two distinct subtasks: mention detection, which involves extracting all relevant mentions from a given text, and mention clustering, which groups these mentions into coreference clusters based on their referential relationships [8].

Recent advancements in coreference resolution have been driven by applying deep neural network systems, particularly transformer architectures such as BERT [9]. These models leverage high-dimensional, multi-scale representations to capture nuanced linguistic patterns and dependencies, significantly improving the accuracy of coreference resolution tasks [10-12]. Furthermore, the versatility of transformer architectures has facilitated the development of end-to-end models, such as seq2seq architectures, which handle both mention detection and clustering jointly [8, 13]. This unified approach streamlines the coreference resolution process and has shown promising results in various domains, including the geological domain [14] and multilingual settings [15].

Pronoun resolution, a particularly challenging subtask within coreference resolution [16], is pivotal for achieving coherent text understanding and generation, with significant implications for tasks such as machine translation. However, this task becomes even more complex for under-resourced languages like Persian, where the scarcity of specialized preprocessing tools and annotated corpora hinders the development of effective models. While most coreference and pronoun resolution systems have primarily focused on English and Chinese, researchers working with less-studied languages, such as Persian or Turkish [17], encounter unique challenges due to the lack of tailored linguistic resources. This paper explicitly targets the Persian language, recognizing that models performing well in other languages may not

✉ **Alireza Talebpour**
Talebpour@sbu.ac.ir

generalize effectively to Persian due to structural differences.

Coreference resolution systems are broadly classified into rule-based and machine-learning paradigms [18]. Early rule-based systems relied on hand-crafted rules for mention clustering [19-24]. Among these, the module-based approach, which employs a series of rule-based modules arranged in descending order of accuracy, has been particularly successful [25-27]. However, while rule-based systems demonstrate simplicity and efficacy in extracting entity-level features [28], they often fall short in pronoun resolution, necessitating sophisticated inference from a combination of linguistic and contextual cues. This complex inferential task is often better suited for machine learning models, particularly those leveraging large-scale pre-trained language models for enhanced contextual understanding.

Machine learning systems, leveraging rich semantic features derived from training data, offer a promising avenue for addressing the complexities of pronoun resolution. While early approaches such as mention-pair and mention-ranking models have demonstrated efficacy, their local focus often needs to improve their ability to make accurate mention merge decisions, especially in long-range dependencies [1]. This limitation has spurred the development of entity-based models, which operate at a cluster level and incorporate broader contextual information. However, entity-based models can encounter computational bottlenecks when processing large documents with extensive coreference chains.

End-to-end models, which jointly handle all stages of coreference resolution, have recently emerged as powerful contenders, demonstrating remarkable performance in specific settings [29, 30]. Nevertheless, these models grapple with challenges such as scaling to large documents, requiring substantial training data, and addressing biases in their predictions [28]. Moreover, despite notable advancements, pronoun resolution remains formidable due to the intricate interplay of contextual information and external knowledge [31]. The effective extraction and integration of external knowledge solely from training data pose a persistent challenge [32].

Recent research has further illuminated the complexities of pronoun resolution. [33] highlight the importance of anisotropic span embeddings in capturing the varied significance of different mention spans while emphasizing the potential negative impact of higher-order inference on model performance. Additionally, the need for fine-tuning coreference models for diverse linguistic styles, such as those found in clinical narratives, has emerged as a critical area of investigation.

The availability of high-quality annotated corpora is essential for developing effective coreference resolution systems [34]. This study leverages two corpora: the established RCDAT [35] and the newly developed Mehr corpus [36]. We propose a hybrid approach to address pronoun resolution challenges, integrating rule-based modules with a machine-learning module that utilizes hand-extracted features, Glove, and Persian BERT embeddings. The entity-centric design, incorporating cluster-level features, significantly enhances system efficiency.

Our hybrid model, incorporating random forest and MLP neural network classifiers, has demonstrated promising results in Persian pronoun resolution. The random forest classifier has emerged as a particularly effective component due to its capacity to model complex interactions among diverse linguistic features and its inherent interpretability [37]. This effectiveness is further amplified by integrating BERT embeddings into the feature set, allowing the model to leverage rich contextual information and enhance its predictive power. The system's architecture is modular and simple, with all components implemented as sequential processing pipelines. This facilitates a streamlined resolution process, efficiently identifying suitable antecedents for each anaphoric pronoun.

The findings of this study have implications for pronoun resolution, particularly for under-resourced languages like Persian. By addressing the unique challenges faced in this domain and introducing a novel hybrid model enriched with BERT embeddings, our research provides insights and practical solutions that can be leveraged to enhance the accuracy and efficiency of pronoun resolution systems across various languages. Developing the Mehr corpus, a comprehensive Persian coreference resource further contributes to advancing Persian NLP research. The evaluation of our proposed system against existing models on both the Mehr and Uppsala test sets offers evidence of its performance, establishing a new benchmark for Persian pronoun resolution.

The remainder of this article is structured as follows: Section 2 analyzes existing coreference resolution systems, encompassing both English and Persian approaches, focusing on hybrid and transformer-based models. Section 3 delves into the landscape of Persian coreference corpora, offering a critical review and highlighting the unique contributions of our newly developed Mehr corpus. Section 4 details our proposed hybrid model for Persian pronoun resolution, elucidating its architecture, the integration of Persian BERT embeddings, and the role of cluster-level features. Section 5 presents an evaluation of the model, outlining the experimental setup, datasets, and metrics employed and discussing the results obtained on both the RCDAT and Mehr corpora. Finally, Sections 6 and 7 make a discussion and conclude the paper by summarizing the key findings, discussing the implications of our work, and proposing avenues for future research.

2- Related Work

This section provides a comprehensive overview of existing coreference resolution systems, encompassing seminal works in English and recent advancements tailored to the Persian language. We will examine the evolution of methodologies, from early rule-based approaches to cutting-edge deep learning techniques, highlighting key contributions and persistent challenges in the field. Additionally, we will delve into the unique linguistic and resource-related obstacles faced in Persian coreference resolution, setting the stage for our proposed hybrid approach.

2-1- Coreference Resolution Systems in English

A seminal work in rule-based coreference resolution is the system proposed by [27] , which utilizes a series of modules organized in a precision-ranked cascade. This approach prioritizes high-precision rules in earlier modules, gradually transitioning to more recall-oriented rules in later modules. This system's influence is evident in its adoption within Stanford University's preprocessing tools and its adaptation for coreference resolution in various languages, including Indonesian [38].

Early machine-learning approaches to coreference resolution primarily utilized the mention-pair model, which employs a binary classifier to assess the coreference relationship between each mention and potential antecedent [2, 39-42]. While simple and widely adopted, this approach is inherently limited by its local focus, considering only pairwise relationships between mentions. [43, 44] offers a broader perspective by evaluating multiple candidates simultaneously. This model allows for a non-anaphoric mention, i.e., not referring to any previous entity. The mention-ranking architecture has been successfully integrated into numerous deep-learning systems [45, 46] demonstrating its effectiveness in capturing more complex coreference patterns.

To address the limitations of local focus in early mention-pair models, entity-based models [47-49] emerged, leveraging cluster-level features for broader contextual analysis. This led to the development of the cluster-ranking model, which combines entity-based and ranking approaches to score potential antecedents at the cluster level. Simultaneously, hybrid systems have been proposed to exploit the strengths of both rule-based and machine-learning paradigms [50]. These hybrid systems often employ dedicated machine-learning modules for different entity types [50] or use dependency trees for enhanced linguistic analysis [28].

A recent trend in coreference resolution is the adoption of end-to-end neural models, which jointly handle mention detection and clustering, eliminating the need for separate preprocessing and feature engineering steps [46]. These models often employ advanced neural architectures such as bidirectional LSTMs [46] or transformers Lai et al. [51] to capture contextual information and model complex linguistic relationships. For example, [46] introduced an end-to-end mention-ranking model using a bidirectional LSTM to compute span embeddings, while [52] further enhanced this approach by incorporating a biaffine attention mechanism to model pairwise mention interactions.

Building upon the success of end-to-end neural models, recent research has explored novel architectures and training strategies to improve performance further. The integration of pre-trained language models like BERT has significantly boosted the ability of these models to capture contextual information, leading to substantial gains in accuracy [30, 53]. Additionally, approaches like entity equalization [54] and graph-based modeling [55] have been proposed to enhance global coherence and model complex coreference structures. Seq2seq models, in particular, have shown promise in coreference resolution due to their ability to generate coreference [8, 56] directly.

Despite these advances, end-to-end coreference models still face several challenges. One major issue is scaling to large documents with extensive coreference chains, often requiring substantial computational resources and training data. Additionally, these models can need help with higher-order inference and the effective integration of external knowledge [33]. To mitigate these issues, researchers have explored techniques like antecedent pruning [29], dual caching mechanisms [57] for efficient processing of long documents, and innovative architectures that balance the need for both local and global information [33]. Table 1 summarizes the key aspects of the mentioned related works in English coreference resolution.

2-2- Coreference Resolution Systems in Persian

In the context of the Persian language, [58] introduced the first pronoun resolution system designed explicitly for Persian. Their system employed the ranking method to determine the antecedent of pronouns in the PCAC-2008 corpus, which the authors themselves developed. The study [59] elaborated on their system's methodological details and features. The features utilized in their system encompassed pronoun features, antecedent features, and relational features. To evaluate the effectiveness of their approach, the authors compared their features with four models: max entropy, max classifier, perceptron, and c4.5. The results section indicated that the c4.5 classifier achieved the highest accuracy, with an F1 score of 44.7.

Table 1: Summary of Key Contributions in Related Works on Coreference Resolution

Article	Approach/Model	Key Contributions
[27]	Rule-Based	Successful rule-based system using modules with decreasing precision and increasing recall.
[50]	Hybrid (ML + Rule-based)	Hybrid system with dedicated machine learning modules for each entity type; explored the use of dependency trees.
[46]	End-to-End (Mention-Ranking)	Bidirectional LSTM for span representation; joint mention detection and clustering.
[52]	End-to-End (Span-Ranking)	Improved span representation with biaffine attention; joint optimization of mention-ranking and coreference-ranking objectives.
[60]	End-to-End (Higher-Order)	Incorporates higher-order inference for improved accuracy; employs antecedent pruning for efficiency.
[51]	End-to-End (Transformer-based)	Integration of pre-trained transformer models (e.g., BERT) to leverage contextualized language representations.
[61]	Neural Network + Discourse Theory	Combined neural networks with centering theory for a discourse-aware approach.
[54]	Entity Equalization	Introduced entity equalization to consider each mention in the context of its entire coreference chain.
[30, 53]	Contextual Text Embeddings	Utilized contextual text embeddings (e.g., BERT, SpanBERT) to enhance the representation of mentions and improve accuracy.
[62]	Graph Model (GT2G)	Proposed a graph-based model for encoding coreference links, generating complete coreference graphs directly.
[63]	BERT-based (Actor-Critic Method)	Achieved state-of-the-art performance on CoNLL-2012 with an actor-critic reinforcement learning framework for joint mention detection and clustering.
[33]	Anisotropic Span Embeddings	Highlights the importance of anisotropic span embeddings and analyzes the potential negative impact of higher-order inference.
[57]	Dual Cache Mechanism	Introduces a dual cache mechanism to improve efficiency in handling long documents with extensive coreference chains.

In their study, [64] introduced a novel pronoun resolution corpus tailored explicitly for the Persian language. Building upon the previous work [58], the authors enhanced the system's efficiency by incorporating a new feature set into their newly developed corpus, and including these additional features led to a remarkable 20% increase in performance. Furthermore, [65] improved the results by applying a series of heuristics to solve the problem of an unbalanced dataset generated in the system [58] and adding semantic information. To evaluate the effectiveness of their approach, the authors examined various classifiers, including maximum entropy, perceptron, SVM, random forest, KNN, and C4.5. The best-reported results were the decision tree, about 75 F1 on the PCAC-2008 corpus.

[66] proposed a rule-based system explicitly designed for Persian pronoun resolution. The system consists of two main components: pre-processing and antecedent selection. By employing five simple rules, the system determines the antecedent of pronouns, considering up to three sentences before the pronoun occurrence. The researchers conducted their experiments using a corpus of five Persian blogs containing 20 pages. A comparative analysis between their system and the one presented in [58] revealed that it achieved higher accuracy and recall rates, indicating improved pronoun resolution performance.

[67] introduced the first coreference resolution system for Persian. The authors developed the Lotus corpus and used 17 features to construct their coreference system. The system was tested on 20% of the Lotus corpus in the evaluation phase using various classifiers, including a decision tree, support vector machine (SVM), and neural network (NN). Among the classifiers, the neural network approach demonstrated the best performance, achieving an accuracy of 39.4 F1 metric.

The previous systems developed for coreference resolution in Persian were primarily based on statistical approaches and utilized limited corpora for their experiments. However, these systems could not form output chains and did not employ widely recognized evaluation metrics such as CoNLL for performance assessment. The research direction in Persian coreference resolution took a significant turn with the development of the RCDAT corpus [35] , which consists of one million tokens. The corpus developers reported 60 conll points by developing the mention-pair model using the SVM algorithm on the RCDAT corpus.

Another notable system that evaluated its results on the RCDAT corpus is the system introduced in [39]. This article employed a fully connected neural network as the coreference classifier. A graph-based algorithm was utilized to construct coreference chains, leading to a significant improvement of 2.2 CoNLL points compared to the previous system. This article employed a fully

connected neural network as the coreference classifier. A graph-based algorithm was utilized to construct coreference chains, significantly improving 2.2 CoNLL points compared to the previous system. In their subsequent paper [40], the authors further enhanced the results by maintaining the general framework of the model while modifying the feature vector and incorporating a semantic feature introduced in [68]. The feature set in this article consisted of three components: heuristic features, word vectors, and semantic features. Coreference chains were formed using a hierarchical clustering algorithm to generate the output. Notably, they reported a remarkable result of 64.54 CoNLL points on the RCDAT corpus, showcasing the effectiveness of their approach and the advancements made in Persian coreference resolution.

3- Coreference Corpora in Persian

The availability of high-quality annotated corpora is a prerequisite for developing and evaluating effective coreference resolution systems. In this study, we leverage two distinct Persian corpora: the established RCDAT corpus and the newly developed Mehr corpus. The RCDAT corpus comprises 1,599 documents and is a valuable resource for evaluating coreference systems in Persian. However, it exhibits limitations due to the absence of annotations for specific linguistic phenomena, such as non-referential pronouns and nested named entities, which are crucial for comprehensive pronoun resolution.

We have developed the Mehr corpus to address these limitations and provide a more comprehensive resource for Persian coreference research. In the following parts of this section, we comprehensively examine the available corpora in Persian. Furthermore, we provide detailed insights into the development process and specifics of the Mehr corpus, shedding light on its construction methodology and characteristics.

3-1- Early Corpora in Persian

The PCAC-2008 corpus [58], a pioneering effort in Persian pronoun resolution, consists of 31 texts extracted from BijanKhan's corpus [69]. However, it is limited in scope, focusing solely on the closest antecedent annotation for a subset of pronouns (2006) and needing more diversity in text genres and topics, potentially hindering its generalizability. In contrast, the Lotus corpus [67] represents the first comprehensive Persian coreference resolution corpus. However, its annotation is limited to coreferent noun phrases within 50 lengthy texts from BijanKhan's corpus, raising concerns about its ability to capture the full spectrum of coreference phenomena in Persian.

3-2- PerCoref Corpus

The PerCoref corpus [70] , comprising 212,646 tokens across 547 documents, represents a notable effort to compile a diverse dataset for Persian coreference resolution. However, its construction methodology, which involves random collection of documents from the web, raises concerns regarding the consistency and potential biases in its annotation. This methodological limitation may hinder the generalizability of findings derived from experiments conducted on this corpus. Nevertheless, PerCoref encompasses a wide range of coreference phenomena, including three reference types (direct, indirect, and none) and six coreference relations (Identity, person/number suffix on the verb, Inferred, Event, Quantifier, and Cross-speech), offering a potentially valuable resource for exploring the complexities of Persian coreference. Table 2 provides detailed statistics regarding the composition of this corpus, and Table 3 presents the statistical distribution of these coreference types within the corpus.

Table 2: PerCoref statistics

Number of documents	547
Number of tokens	212646
Number of sentences	6688
Number of paragraphs	4449
Average paragraph per document	8.13
Average sentence per document	1.5
Average token per document	31.79

Table 3: PerCoref corpus coreference relations

Coreference Relations	**Count**	
Identity	11812	58.4
person/number suffix on verb	3021	14.9
Inferred	2005	9.9
Event	1509	7.4
Quantifier	968	4.7
Cross-speech	901	4.4

3-3- RCDAT Corpus

The RCDAT corpus [35] , comprising one million words annotated in the CoNLL format, is a substantial resource for Persian coreference resolution. It includes gold standard annotations for coreference chains, named entities, and animacy and covers a various political, sports, and cultural texts from various news sources. However, its annotation scheme lacks coverage of important linguistic phenomena pertinent to pronoun resolution, such as non-referential pronouns and nested named entities. This limitation constrains its utility for comprehensive pronoun resolution tasks, particularly those that require a deeper understanding of referential ambiguities and complex syntactic structures. Table 4 provides detailed statistical information about the RCDAT corpus.

Table 4: RCDAT's statistical information

	Count
Number of documents	1599
Number of tokens	1052637
Number of labeled tokens	408138
Number of entities	86960
Number of pronouns	2687
Number of proper nouns	3412
Number of animate entities	4266

Each line of the RCDAT corpus provides detailed information about individual tokens, including the document name, sentence number, token itself, part-of-speech tags (16 tags), named entity tags (both 13 and 3 granularities), the token's root form, original token form, coreference chain index (gold standard), coreference chain number (gold standard), animacy, phrase type, and a more fine-grained set of part-of-speech tags (100 tags).

3-4- Mehr Corpus

To address the limitations of existing Persian corpora, the authors developed the Mehr corpus [36], a comprehensive resource designed explicitly for Persian coreference resolution. This corpus comprises 400 texts randomly extracted from the Mehr news agency, ensuring diverse topics and linguistic styles. Annotation was meticulously performed by two expert annotators using WebAnno [71] , a collaborative annotation tool that facilitates high-quality and consistent annotations. WebAnno also automatically generates the CoNLL file for the labeled chains, streamlining the data preparation process for subsequent model training and evaluation. The Mehr corpus is divided into distinct training, development, and testing sets, following established best practices for model development and evaluation. The test set, comprising approximately 13,000 tokens, is specifically designed to align with the size of the MUC7 corpus test set [72], enabling direct comparison and standardized evaluation against other coreference resolution systems.
Fig1 visually represents the annotation process within the WebAnno environment, illustrating the user interface and annotation workflow. Table 5 presents a comprehensive summary of the statistical characteristics of the Mehr corpus, including the number of tokens, sentences, and documents in each split. This detailed information allows a thorough understanding of the corpus composition and its suitability for various coreference resolution tasks.

Table 5: Statistical characteristics of Mehr's corpus

Document	**Document Counts**	**Sentence Counts**	**Token Counts**
Train and Development	357	3189	158047
Test	43	286	13214
Total	400	3475	171261

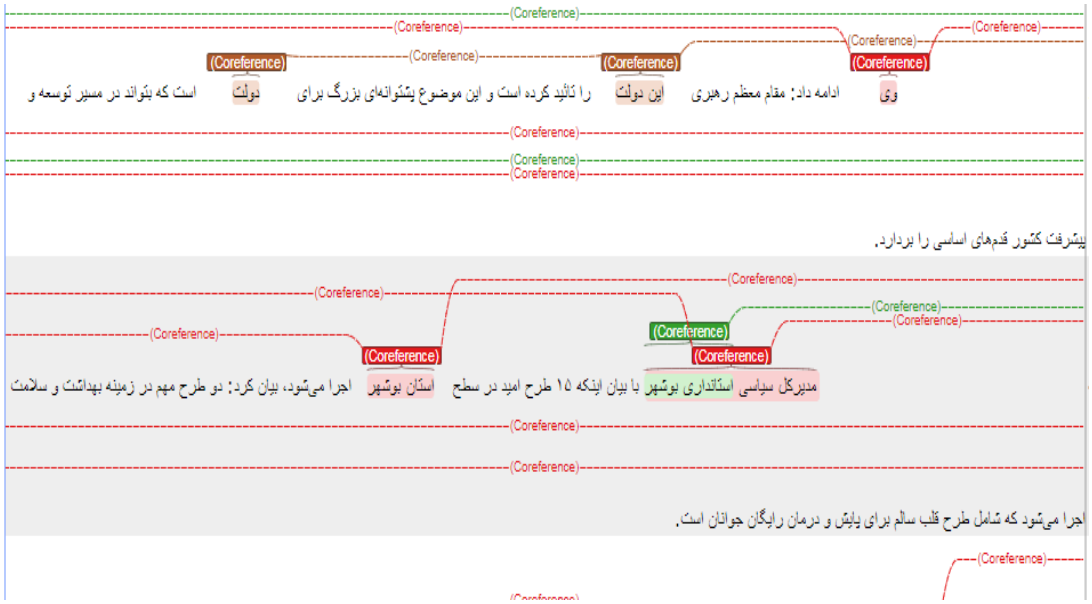


Fig1. A frame of WebAnno annotation tool

The Mehr corpus employs a dual-labeling strategy, distinguishing between "gold labels" and "system labels".

Gold labels:

In the Mehr corpus, the coreference labels for pronouns, noun phrases, named entities, and nested named entities are assigned through manual annotation. Only the mentions of at least one coreference chain are labeled in the corpus. It is worth noting that some non-referential pronouns are labeled among the pronouns. In Persian, some pronouns do not have a direct reference in the text. In fact, in Persian, some pronouns do not refer to anything. Non-referential pronouns are also labeled in the Mehr corpus. Some pronouns, such as nobody and others, which do not have any reference in the text, are not labeled.

In Persian, some pronouns have an indirect reference. These pronouns refer to the speaker of a quote or an utterance, which may have been omitted in the document. In the Mehr corpus, these types of pronouns with indirect reference are also labeled and included in the coreference annotations. In the Mehr corpus, the focus is on labeling pronominal singletons with coreference relations. Specifically, only pronominal mentions with no explicit antecedents within the corpus are labeled. The coreference relation type assigned to these pronominal singletons is the identity relation, indicating that they refer to the same entity. The following is an example of each coreference relation in the Mehr corpus.

- **Coreference Relations**

(محمد) به مدرسه آمد. بچه ها از دیدن (او) تعجب کردند.

(Muhammad) came to school. The children were surprised to see (him).

(ابن سینا) دانشمند ایرانی است. (این دانشمند برجسته) در همدان مدفون است.

IbnSina is an Iranian scientist. (This outstanding scientist) is buried in Hamadan.

- **Indirect Reference**

(ما) باید به تیم ملی خود افتخار کنیم.

(We) should be proud of our national team.

- **Non-Referential Pronouns**

(این)، مساله مهمی است.

(It) is an important issue.

System labels:

Text processing tools in Persian automatically generate labels for various linguistic elements. The accuracy of these automatically generated labels may vary compared to manual labels and relies on the precision of the preprocessing tools employed. The Mehr corpus's automatic labeling encompasses elements such as named entities, noun phrases, POS tags, and syntactic groups. This corpus uses the Hazm preprocessing tool [1], the comprehensive platform of intelligent content analysis [2], and the FarsiYar text preprocessing tool [3] .Table 6 shows the accuracy of preprocessing tools used in this corpus.

Including gold and system labels in the Mehr corpus aims to provide a comprehensive and multifaceted resource for advancing research in Persian coreference resolution. The manual annotations ensure high-quality training and evaluation data, while the automatic labels offer complementary information for further linguistic exploration.

[1] https://www.sobhe.ir/hazm/
[2] http://catotalservices.com/UIHome/Index.aspx
[3] https://text-mining.ir/

Table 6: Accuracy of preprocessing tools used in this paper

Persian Tools	Accuracy
Pos tagger	97.1
Shallow parser (Chunker)	89.9
Constituency parser	85.2
Named-entity tagger	89
Animacy recognition	91
Pronoun recognition	96

4- Proposed Method

In light of previous research [40] established a benchmark for Persian coreference resolution by demonstrating the superior performance of their system, which implemented state-of-the-art English models on the Persian RCDAT corpus. Building upon this foundation, the present work aims to advance the state of the art by introducing a novel hybrid model that synergistically integrates rule-based and machine-learning components. This innovative approach seeks to leverage the strengths of both paradigms, overcoming the limitations of purely rule-based or machine-learning systems and ultimately achieving superior performance in Persian coreference resolution.

This section establishes a foundation for our investigation by introducing two baseline models for Persian pronoun resolution. The first is a rule-based module-based system inspired by the work of [27], adapted and tailored for the nuances of Persian grammar and syntax. The second baseline is a mention-pair machine learning model that leverages hand-crafted features to capture pronoun, antecedent, and relational information. To further advance the state of the art, we propose a novel hybrid entity-centric model that integrates these two baselines and incorporates additional cluster-level features, along with GLOVE and Persian BERT embeddings, to enrich contextual representation. Our empirical evaluation demonstrates that this hybrid approach significantly outperforms existing Persian systems, highlighting the potential of combining rule-based and machine-learning paradigms for this challenging task.

4-1- Persian Module-Based Baseline Model

Inspired by the widely adopted architecture of [27], our Persian module-based baseline system leverages a precision-ranked cascade of rule-based modules to resolve coreference chains incrementally. Each module operates by merging a mention into an existing partial cluster or deferring antecedent identification to subsequent modules. This approach facilitates incorporating local (mention-level) and global (cluster-level) information in the resolution process.

The system consists of eight modules specifically adapted for Persian, acknowledging the linguistic nuances and structural differences compared to English. Before module application, input documents undergo preprocessing using existing NLP tools to identify mentioned boundaries and extract relevant features such as number and animacy. These identified mentions are then sequentially processed by the rule-based modules, each contributing to the construction of coreference chains hierarchically.

This modular architecture enables the seamless integration of additional rule-based components and ensures transparency and interpretability in the coreference resolution process. While effective for capturing entity-level features, this system's rule-based nature may pose challenges in handling more complex referential relationships, particularly those involving pronouns that require deeper semantic understanding. However, this baseline is a valuable starting point for our subsequent exploration of hybrid models that integrate machine-learning techniques for enhanced pronoun resolution.

4-1-1- Model Architecture

The architecture of mention selection within our rule-based system closely follows the approach proposed by[26], with a critical adaptation to accommodate the right-to-left reading direction of Persian text. Since partial coreference chains are incrementally formed within each module, selecting representative mentions becomes crucial. Our system prioritizes the examination of the first mention within each partial chain, as this mention typically provides a more explicit entity definition and is less likely to be a pronoun or contain morphological ambiguities [27]. This selection strategy also tends to result in a smaller pool of potential antecedents, leading to increased efficiency.

To ensure that the most relevant antecedents are considered, the first mentions within each sentence are sorted from right to left, reflecting the natural reading order of Persian. If no suitable antecedent is found within the current sentence, the algorithm examines previous sentences in the document, again considering mentions from right to left. This systematic approach maximizes the likelihood of identifying the correct antecedent while minimizing computational overhead. The algorithm for the mentioned selection in the current module is outlined as follows.

Selecting Mentions within the Current Module:

Selecting mentions within the current module involves considering the partial chains generated by the previous modules. Only the first mention within each partial chain is selected for examination in the current module. For

instance, let us assume that the partial chains $[m_1^1.m_2^2.m_3^2.m_4^3.m_5^1.m_6^2]$ have been constructed for the current document. The lower index denotes the mention number, while the upper index represents the cluster number. In the current module, the model focuses on finding the antecedent for m_2^2 and m_4^3 (m_1^1 is not checked as it is located at the beginning of the document). This selection is based on the observation that first mentions within partial chains provide better entity definitions and have a lower probability of being pronouns or containing prefixes and suffixes[27]. Additionally, first mentions are typically closer to the beginning of the text, resulting in a smaller pool of potential antecedents to consider.

Antecedent Selection for a Current Mention: Each mention in the current module may be merged with a previous partial chain, or its antecedent may be assigned in the next module. For each mention in the current module, the mentions in the current sentence are sorted from right to left and checked in order. The document sentences are traversed from left to right if there is no answer in the current sentence for the current mention. Then, in each sentence, the mentions are sorted from right to left to reach the answer. Sorting the list of candidates is important because the algorithm stops at the first confluence. Weak sorting causes the accuracy of the system to decrease. A sentence threshold window should be set in each module so the algorithm does not check more than that threshold limit.

Sharing Features in the Entity-Centric Model: When constructing a partial entity that consists of one or more mentions, the system performs a union operation on the features of those mentions, such as number, gender, and animacy. The resulting feature values are then shared among all the mentions within that entity. In cases where the feature values differ among the mentions of the entity, all the values are preserved. For example, suppose the number detection module identifies the mention "group of students" as singular and "five students" as plural. In that case, the resulting partial entity will have singular and plural values for the number feature. This allows for including singular and plural pronouns in the pronoun resolution module when referring to this partial entity.

4-1-2- Model Modules

Our rule-based system employs eight distinct modules, each designed to address specific coreference scenarios in Persian. These modules operate sequentially, each contributing to the incremental construction of coreference chains. The order of the modules is carefully determined, prioritizing high-precision rules in earlier stages and gradually transitioning to more relaxed rules for improved recall. This hierarchical approach ensures a systematic and efficient resolution process while accounting for the varying complexities of different coreference phenomena. We will now detail the functionality of each module within this framework.

Module 1 - Speaker Module: The initial module in our system focuses on resolving pronouns within quotations, a common linguistic construct in news articles and other written discourse. The speaker is typically explicitly mentioned in conversational texts, facilitating straightforward pronoun resolution. However, the speaker may be implicitly or indirectly referenced in non-conversational texts, such as those found in the Mehr and RCDAT corpora. To address this, Module 1 employs a set of rules designed to identify quoted verbs (e.g., "said," "stated," "announced") and their associated subjects, which often correspond to the speaker of the quotation. Once the speaker is identified, the module resolves pronouns within the quotation to this antecedent, establishing a coreference link based on the discourse context.

Module 2 – Exact String Match: Module 2, the Exact String Match module, is a high-precision component that operates on the principle of string identity. In this module, if two mentions match precisely, including any prefixes or suffixes, they are deemed coreferent and merged into a single chain. This approach is computationally efficient and particularly effective in resolving mentions with identical surface forms, such as repeated proper names or nominal mentions. While exact string matching does not consider deeper linguistic or semantic information, its high precision makes it a valuable initial step in the coreference resolution pipeline, resolving straightforward cases before resorting to more complex analysis in subsequent modules.

Module 3 – Strict Head Match: Building upon the string-based matching of Module 2 and Module 3, the Strict Head Match module applies more stringent criteria for resolving coreferences based on head noun agreement. While mentioning sharing the same head noun as a potential candidate for coreference, this module acknowledges that head noun overlap alone is insufficient to guarantee referential identity. For instance, "Tehran University" and "Washington University" share the head noun "University" but refer to distinct entities. Stricter conditions for merging mentions based on head noun matching are imposed to mitigate such errors. The head noun of the current mention must match the head noun of a previous mention and be a complete substring of that mention. For example, the mentions "Tehran Court" and "Tehran High Court" would be considered coreferent as the former is a substring of the latter. This approach ensures higher precision in coreference resolution by considering lexical and structural cues.

Module 4 – Proper Name Match: Module 4 is dedicated to resolving coreference involving proper names. This module targets mentions that share the same proper noun head, such as "Beckham" and "David Beckham," to merge them into a single coreference chain. To ensure accuracy, certain conditions are imposed to prevent erroneous merges. For instance, the module may consider the relative positions of the mentions within the text or the presence of additional modifiers that could differentiate the referents. This approach helps refine the coreference resolution process by addressing the unique challenges posed by proper name mentions in Persian.

Module 5 – Location Match: Module 5 focuses on the resolution of location-based coreference. It specifically targets mentions tagged as "LOC" (location) by the named entity recognition (NER) component of the pre-processing pipeline. In this module, a substring matching heuristic is applied to determine coreference. If one location mention is a complete substring of another, they are considered to refer to the same geographical entity and are merged into a single chain. For instance, the mentions "Tehran" and "Tehran City" would be linked under this rule. This module aims to capture coreference relationships that arise from the hierarchical nature of location names, where more specific locations are often embedded within broader geographical entities. By identifying and merging such mentions, Module 5 enhances the accuracy of coreference resolution for location-related entities in Persian texts.

Module 6 – Title Match: Module 6 addresses coreference relationships involving titles, appositives, and descriptive phrases associated with person names. In Persian news texts, it is common for individuals to be referred to by their titles or appositives, either preceding or following their names. For example, a sentence might mention "President Hassan Rouhani" followed by "Rouhani" in a later sentence, or vice-versa. This module leverages simple pattern-matching rules to identify and merge such pairs into a single coreference chain. By recognizing the relationship between names and their associated titles or descriptions, Module 6 enhances the accuracy and completeness of coreference resolution for person mentions in Persian texts.

Module 7: Demonstrative Phrase Match: Module 7 addresses coreference resolution involving both demonstrative phrases and pronouns. In Persian, demonstratives often play a crucial role in establishing coreference relationships, as they explicitly point to previously mentioned entities. This module identifies mentions containing demonstrative expressions (e.g., " این نمایشگاه" meaning "this exhibition") and attempts to link them with their corresponding antecedents based on semantic and syntactic cues. In addition to demonstratives, this module handles pronoun resolution for pronouns not resolved by earlier modules. It uses syntactic and semantic features to identify the most likely antecedent for each pronoun within a defined context window. This dual functionality enhances the system's ability to capture a broader range of coreference phenomena in Persian text.

Module 8: Pronoun Resolution: In our proposed module-based system, the final module is dedicated to resolving unresolved pronoun mentions after applying the previous seven modules. This module leverages a feature-matching approach to link each pronoun to one of the existing partial chains, thereby identifying its most likely antecedent. Table 7 compares the features used in the proposed system and the system introduced in reference [27].

Table 7: Features related to the pronoun module in the system [27] and the presented system

Feature Name	Proposed System	English Module-based System
Number	1- Static list for pronouns 2- NER labels 3- POS labels	1- Static list for pronouns 2- NER labels 3- POS labels 4- Static dictionary from [73]
Gender	1- Static lexicon 2- Static pronouns	1- Static lexicon from [73]
Person	-	1- Use for only pronouns
Animacy	1- Static list for pronouns 2- NER labels 3- Static list for titles 4- Static list for Persian names 5- Entity label from Persian corpus [۷۵]	1- Static list for pronouns 2- NER labels 3- Dictionary extracted from WEB [74]
Distance	Less than 4 sentence	Less than 4 sentence

4-2- Persian Mention-Pair Baseline Model

Our second baseline model employs a mention-pair architecture, a classical coreference resolution approach focusing on pairwise comparisons between potential mention-antecedent pairs. This model utilizes a suite of hand-crafted features meticulously designed to capture the linguistic nuances of the Persian language. These features encompass various aspects of pronouns (e.g., type, grammatical role), their potential antecedents (e.g., entity type, syntactic position), and the relationship between them (e.g., distance, agreement). We introduce a distance window hyper-parameter to enhance the model's generalization capabilities, which restricts the maximum distance between a pronoun and its potential antecedent. This helps mitigate the risk of erroneous long-range coreference links.

During training, we employ a strategy to address class imbalance, a common issue in coreference datasets where negative examples (non-coreferent pairs) significantly outnumber positive examples (coreferent pairs). We create a positive example for each pronoun by pairing it with its nearest true antecedent. To balance the dataset, we generate negative examples by including all intervening mentions between the pronoun and its true antecedent [41].

A comprehensive feature vector encompassing mention-level and cluster-level attributes (Table 8) is constructed for each mention-antecedent pair. The features are mention-level (M) and cluster-level (E). The "+" sign indicates features exclusive to the hybrid model presented in this article, while the mention-pair system incorporates features without the "+" sign.

These features include lexical, syntactic, semantic, and positional information, providing a rich representation for the machine learning classifier. Through rigorous experimentation involving five repetitions of 10-fold cross-validation on the training data, we evaluated the performance of various classifiers, including decision trees, random forests, multi-layer perceptron, and support vector machines. The results presented in Section 5 demonstrate that the random forest classifier consistently outperforms the others for this mention-pair baseline model.

Table 8: The features used in the proposed hybrid models and the mention-pair model

Pronoun Features		
row	Description	M/E
1	Is it a personal pronoun?	M
2	Is it a demonstrative pronoun?	M
3	Is it a reflexive pronoun?	M
4	Is it a third-person pronoun?	M
5	Is it a speech pronoun?	M
6	Pos label of the first token on the left side of the pronoun	M
7	Pos label of the second token on the left side of the pronoun	M
8	Pos label of the third token on the left side of the pronoun	M
9	Pos label of the first token on the right side of the pronoun	M
10	Pos label of the second token on the right side of the pronoun	M
11	Pos label of the third token on the right side of the pronoun	M
12	Is pronoun subject?	M
13	Is pronoun object?	M
14	Pronoun number	M
15	Pronoun animacy	M +
16	Pronoun person	M +
Antecedent Features		
17	How many tokens does the antecedent have?	M
18	Is the antecedent a pronoun?	M
19	Is the antecedent a demonstrative phrase?	M
20	Pos label of the first token on the left side of the antecedent	M
21	Pos label of the second token on the left side of the antecedent	M
22	Pos label of the third token on the left side of the antecedent	M
23	Pos label of the first token on the right side of the antecedent	M
24	Pos label of the second token on the right side of the antecedent	M
25	Pos label of the third token on the right side of the antecedent	M
26	Antecedent number	M
27	Is Antecedent subject?	M

28	Is Antecedent object?	M
29	How many mentions are there in the antecedent chain?	E+
30	Is the antecedent a reflexive pronoun?	M+
31	Antecedent type (pronoun, proper noun, common noun)	M+
32	The sentence Index of the first mention in the antecedent chain	E+
33	Antecedent animacy	M+
34	Antecedent person	M+
35	Named-entity label	M+
36	Antecedent chain number	E+
37	Antecedent chain animacy	E+
Relational Features		
38	Sentence distance	M
39	Token distance	M
40	Number agreement	M
41	Subject agreement	M
42	Object agreement	M
43	String match	M
44	Is the distance between the pronoun and antecedent candidate less than three?	M
45	Are the pronoun and the antecedent candidate in the same sentence?	M
46	What is the minimum sentence distance between the pronoun and the antecedent candidate chain?	E+
47	Is the antecedent string token count more than the pronoun string?	M+
48	Animacy agreement	M+
49	Person agreement	M+
50	If a pronoun is an object and the antecedent is a subject, both are in the same sentence.	M+
51	If a pronoun is reflexive and the antecedent candidate is a subject, both are in the same sentence.	M+
Word Embedding Features		
52	Euclidean distance of two mention heads	M+
53	Euclidean distance between the mean of two mentions	M+
54	The Euclidean distance between the antecedent candidate and the sentence in which the pronoun is placed.	M+
-	word embedding for pronoun	M+
-	word embedding for antecedent	M+

4-3- The Proposed Hybrid Model

Building upon the rule-based and machine-learning baselines presented in previous sections, we introduce a novel hybrid model for Persian pronoun resolution that combines the strengths of both paradigms. This hybrid system leverages the efficiency and interpretability of rule-based approaches, exemplified by the seven high-precision modules described in Section 4.1 while harnessing the power of machine learning to model complex linguistic relationships and contextual dependencies. The final module in our pipeline, designed explicitly for pronoun resolution, employs a random forest classifier that has been shown to handle high-dimensional feature spaces and provide interpretable results effectively.

The model operates entity-centric, incrementally gathering information about coreference clusters as the text is processed. As depicted in

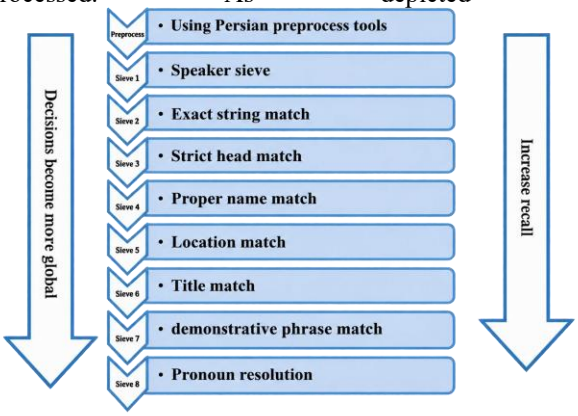


Fig 2 , this incremental approach allows the model to incorporate cluster-level features, such as the number of mentions in a cluster and the animacy of the cluster's referent, into the decision-making process for pronoun resolution. By utilizing both hand-crafted features and contextualized word embeddings derived from Persian BERT and GloVe, the machine-learning module effectively learns to identify the most likely antecedent for each pronoun, considering both local and global contextual information.

Our hybrid model offers several key advantages over existing approaches. By incorporating rule-based modules, we can quickly leverage linguistic knowledge and

heuristics to resolve unambiguous coreferences. This reduces the burden on the machine-learning module, allowing it to focus on more challenging cases where contextual information and complex linguistic patterns are critical. Additionally, using a random forest classifier ensures computational efficiency and interpretability. The subsequent sections will delve deeper into the model's training and testing procedures and a comprehensive evaluation of the Mehr and RCDAT corpora, demonstrating its superior performance compared to previous Persian coreference resolution systems.

Feature Vector: A key component of our hybrid model is carefully engineering features that capture the multifaceted nature of pronoun resolution. Our approach combines hand-crafted features, specifically designed to capture the linguistic nuances of Persian, with rich contextualized representations derived from both GloVe and Persian BERT embeddings. These hand-crafted features encompass a wide range of attributes about pronouns (e.g., type, grammatical role), their potential antecedents (e.g., entity type, syntactic position), and their relationships (e.g., distance, agreement). By integrating GloVe embeddings, we capture semantic and syntactic relationships between words, while Persian BERT embeddings provide a deeper understanding of contextual nuances. This multi-faceted approach to feature engineering enhances the model's ability to disambiguate pronoun references and make informed predictions.

Table 8 provides a comprehensive overview of the 54 features employed in our hybrid model, encompassing mention-level and cluster-level attributes. This diverse set of features, combined with the power of contextual embeddings, enables the random forest classifier to effectively model the complex interactions between pronouns and their potential antecedents, leading to superior performance in Persian pronoun resolution. This comprehensive feature engineering approach, coupled with the hybrid model architecture, contributes to the superior performance of our system on the Persian coreference resolution task.

Training: Our entity-centric approach aims to establish coreference links between pronouns and their potential antecedents by utilizing a machine learning module trained on pre-existing coreference chains. The training process generates positive examples by associating each pronoun with its closest actual antecedent and negative examples by incorporating all intervening mentions within the specified distance window. This strategy aims to mitigate the class imbalance issue commonly encountered in coreference datasets.

Crucially, our model incorporates cluster-level features, such as the number of mentions in a cluster and the animacy of the cluster's referent, to provide a richer contextual representation for pronoun resolution. We evaluate the performance of various classifiers, including random forests, multi-layer perceptron, support vector machines, and gradient boosting trees, on two distinct Persian corpora: the newly developed Mehr corpus and the established RCDAT corpus. The results of this comparative analysis, including the optimal hyperparameters for the best-performing classifier, are presented in Section 5.

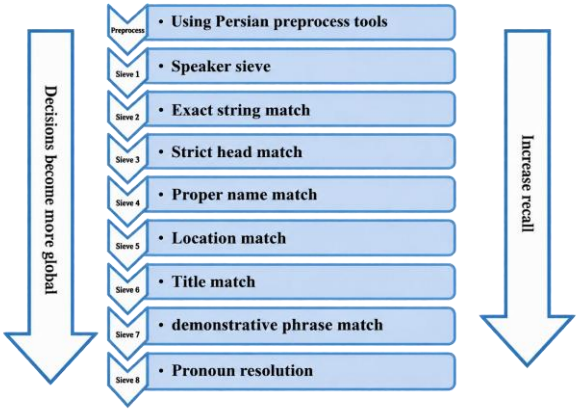


Fig 2. The introduced hybrid architecture.

Test: Our entity-centric hybrid model is evaluated through a systematic testing procedure on unseen data. The process begins with the application of the rule-based modules, generating partial coreference chains based on pre-defined linguistic rules and heuristics. For unresolved pronouns, the machine-learning module, powered by a random forest classifier enhanced with BERT embeddings, assesses potential coreference links with existing partial chains within a predefined sentence window. The random forest classifier assigns scores to each candidate antecedent, selecting the highest-scoring candidate that exceeds a predetermined threshold. If no candidate meets this criterion, the pronoun is classified as non-anaphoric.

5- Experiments

To rigorously evaluate the effectiveness of our proposed hybrid model, we conducted experiments on two distinct Persian corpora: the newly developed Mehr corpus and the established RCDAT corpus. We followed a standard train-test split for the Mehr corpus, utilizing the designated training documents for model development and the corresponding test documents for evaluation. In the case of the RCDAT corpus, due to the absence of a predefined test set, we adopted the methodology employed in previous Persian coreference resolution studies, utilizing the Uppsala test documents [76] for evaluation. This corpus, comprising 415 sentences and 16,274 tokens, offers diverse narratives for assessing the model's performance

on unseen data. It is important to note that our evaluation of the RCDAT corpus focuses solely on pronouns that are part of an existing coreference chain, excluding non-referential pronouns and pronouns initiating new chains, as these are not annotated in the RCDAT corpus. This decision ensures consistency with previous work and directly compares our results with existing systems.

This paper focuses on the specific challenge of resolving detached pronouns in Persian. Detached pronouns in Persian include personal, demonstrative, and reflexive pronouns, each contributing unique complexities to the coreference resolution task. Compared to English, pronoun resolution in Persian poses additional challenges due to the language's distinct linguistic properties.

One prominent challenge stems from Persian's pro-drop or null-subject nature, which allows for omitting nominal subjects in sentences. This phenomenon often obscures the intended referent of a pronoun, making antecedent identification more difficult. Additionally, unlike English, third-person pronouns in Persian lack inherent gender marking, rendering gender an unreliable cue for resolution. Furthermore, Persian's complex system of politeness and formality introduces further ambiguity. Plural pronouns may refer to singular entities out of respect, while singular pronouns may address plural entities in formal contexts. These intricate sociolinguistic nuances further underscore the necessity for specialized models and techniques to navigate Persian pronoun resolution's intricacies effectively.

5-1- Experiments on MEHR Corpus

This section comprehensively evaluates our proposed hybrid model for Persian pronoun resolution on the Mehr corpus. Table 9 summarizes the results of our experiments, directly comparing various baseline systems and previously established Persian coreference resolution approaches. The reported performance metrics reflect the optimal model configurations identified during our experimentation. We begin by establishing baseline results using a rule-based system and a mention-pair model (Lines 1-2), followed by a comparison with the current state-of-the-art Persian system (Lines 3-4), as reported by [40], which notably outperformed contemporary English systems on the Persian RCDAT corpus.

Using a random forest classifier, the proposed hybrid system demonstrates promising results for Persian pronoun resolution across various settings. The model achieves the highest F1 score (66.8) when utilizing gold cluster-based features and BERT embeddings, underscoring the importance of accurate annotations and contextualized language representations for this task. The system maintains competitive performance even with automatically derived cluster features, particularly when combined with either GloVe or BERT embeddings (F1 scores of 64.75 and 65.15, respectively). These results highlight the effectiveness of our hybrid approach and the potential for further improvement by refining automatic feature extraction methods.

Table 9: The reported results of the presented system on Mehr corpus compared to previous works and various settings

	System	Using Rule-Based Modules	GloVe/BERT	Cluster-Based Features	Precision	Recall	F1
1	Rule-based baseline	+	-	-	36.4	35.6	36
2	Mention-pair baseline (random forest)	-	-	system	60	58.82	59.4
3	[40] (random forest)	-	GloVe	system	55.08	54.54	55.08
4	(MLP) [40]	-	GloVe	system	60.47	59.88	60.47
	Proposed System with Different Settings (Random Forest)						
5		+	GloVe	gold	66.56	66.27	66.41
6		+	-	system	64.2	62.94	63.56
7		+	GloVe	system	65.4	64.11	64.75
8		-	-	gold	64.28	63.01	63.64
9		-	GloVe	gold	66.28	64.98	65.62
10		+	BERT	gold	67.2	66.4	66.8
11		+	BERT	system	65.9	64.42	**65.15**

	System	Using Rule-Based Modules	GloVe/BERT	Cluster-Based Features	Precision	Recall	F1
	Proposed System with Different Settings (MLP)						
12		+	GloVe	gold	64.2	62.97	63.56
13		+	GloVe	system	60.8	59.64	60.2
14		+	BERT	gold	64.9	63.81	64.35
15		+	BERT	system	61.3	63.04	**62.16**
16	Changing the modules' order of system line 7	+	GloVe	system	64.9	63.9	64.4
17	Using logistic regression model instead of random forest in system line 7	+	GloVe	system	60.42	58.44	59.41

5-1-1- Mehr Corpus Result Analysis

Line 11 of Table 9 demonstrates the results of the proposed system on the Mehr corpus. This line indicates that the proposed system outperforms the previous best Persian system, increasing the 4.68 F1 score. This highlights the effectiveness of implementing a hybrid architecture using cluster-level features and leveraging the simplicity of the module-based model. Below is an analysis of the results presented in Table 9.

Hybrid Model Superiority: The hybrid model consistently outperforms both the rule-based baseline (Line 1) and the mention-pair baseline (Line 2), underscoring the benefits of integrating linguistic knowledge with machine learning. Notably, the highest F1 score of 66.8 is achieved when combining the hybrid architecture with gold cluster features and BERT embeddings (Line 10), showcasing the synergistic effect of leveraging both rule-based insights and contextualized representations for enhanced pronoun resolution accuracy.

Impact of BERT and GloVe Embeddings: Including BERT embeddings consistently improves performance over models relying solely on GloVe or no embeddings. Combining both embeddings further elevates results, suggesting a complementary relationship in capturing different aspects of linguistic information. BERT's ability to model contextual nuances is particularly evident in its superior performance compared to GloVe in configurations with gold cluster features (Lines 10 vs. 5).

Significance of Cluster-Level Features: The consistent improvement observed when using gold cluster-level features (e.g., Lines 5, 8, 9, 10, 12, 14) emphasizes their importance in providing a richer representation of coreference chains, enabling the model to make more informed decisions about pronoun resolution. This highlights the need to develop accurate automatic feature extraction methods to bridge the performance gap between gold and system-generated features.

Rule-Based Modules Enhance Performance: While the impact of rule-based modules is less pronounced when using BERT embeddings, they still contribute to improved performance, particularly in recall. This suggests that these modules effectively capture certain coreference relationships that the machine learning component alone cannot fully address. Notably, removing rule-based modules in configurations using gold features (Lines 8 and 9) decreases efficiency, indicating their value in resolving specific linguistic patterns.

Classifier Performance: The random forest classifier consistently outperforms the MLP classifier across comparable settings, demonstrating its superior ability to model complex interactions among the diverse features utilized in our hybrid system. This aligns with the established strengths of random forests in handling high-dimensional data and providing interpretable models.

Model Robustness and Adaptability: The hybrid model's performance remains relatively stable across different configurations, indicating its robustness to feature extraction and module order variations. Additionally, the slight decrease in performance observed when altering the module order (Line 16) suggests that the proposed architecture is flexible and can be adapted to different pipelines without significantly sacrificing accuracy.

Module Order Independence: Changing the order of the rule-based modules has a minimal impact on performance, indicating the robustness and flexibility of the hybrid architecture.

Superiority of Non-Linear Models: The significant drop in performance when using logistic regression (a linear model) instead of random forest (a non-linear model)

highlights the importance of capturing non-linear relationships in the feature space for accurate pronoun resolution.

Challenges in System-Generated Features: The performance gap between systems using gold and system-generated features underscores the need for further research on improving automatic feature extraction methods in Persian NLP.

Potential for Future Improvement: While the hybrid model with BERT embeddings shows promising results, there remains room for further improvement, particularly in refining automatic feature extraction and exploring alternative neural architectures.

These analyses highlight the strengths and improvements of the proposed system, demonstrating the importance of the hybrid architecture, cluster-level features, rule-based modules, and the choice of machine learning algorithms in achieving efficient pronoun resolution in Persian. Table 10 summarizes the main observations and their corresponding analyses, providing a concise overview of the key findings from the experiment results.

Table 10: Analysis of Experimental Results for Pronoun Resolution in the Mehr Corpus

Observation	Analysis
Rule-based baseline (Line 1) exhibits low accuracy.	Rule-based modules alone are insufficient for accurate Persian pronoun resolution, highlighting the need for machine learning integration
Mention-pair baseline (Line 2) shows significant improvement over rule-based baseline.	Random forest effectively models complex feature interactions in pronoun resolution, surpassing rule-based approaches.
Hybrid model with system-based features and BERT (Line 11) achieves the highest F1 score.	This configuration demonstrates the synergistic effect of combining linguistic rules, machine learning, and contextual embeddings (GloVe or BERT).
Hybrid models with gold cluster features (Lines 5, 8, 9, 10, 12, 14) outperform those with system features.	Gold-standard annotations for cluster features yield superior results, highlighting the need for high-quality training data and accurate feature extraction.
Removing rule-based modules (Lines 8 and 9) generally decreases performance.	The hybrid architecture is flexible and robust to variations in module sequencing, maintaining consistent performance.
Random forest consistently outperforms MLP (Lines 7, 11, 13, 15 vs. lines 6, 10, 12, 14).	Random forest's ability to model non-linear relationships in high-dimensional feature spaces makes it more suitable for this task than MLP.
Logistic regression underperforms compared to random forest (Line 17).	The superior performance of random forest emphasizes the importance of capturing non-linear patterns in pronoun resolution.
Including GloVe embeddings improves performance in both automatic and gold models.	Word embeddings provide valuable semantic and syntactic information, enhancing the model's ability to disambiguate pronoun references.
BERT embeddings consistently outperform GloVe embeddings when combined with gold cluster features (Lines 10 & 14).	BERT's contextualized representations are particularly effective in capturing nuanced linguistic relationships in Persian pronoun resolution.

5-1-2- Hyper-Parameter Settings for Implemented Systems on Mehr Corpus

The performance of machine learning models is heavily influenced by the selection of appropriate hyperparameters. In this section, we detail the hyper-parameter tuning process employed for our mention-pair baseline and hybrid models on the Mehr corpus.

Mention-pair baseline :A grid search was conducted to identify the optimal hyper-parameter configuration for the random forest classifier. The hyper-parameter space, as outlined in Table 11 included various values for maximum depth, the number of estimators (trees), and the splitting criterion.

Table 11: Random Forest grid-search hyper-parameter values

Hyper-parameter	Values
Max depth	None,10,20,30,40,50,60,70,80,90,100
Number of estimators	100,200,500,1000
Criterion	Gini, Entropy

The grid search revealed that the optimal configuration for the mention-pair baseline model on the Mehr corpus was:

- **Max Depth:** 1000
- **Number of Estimators:** 20
- **Criterion:** Gini

To evaluate the model's robustness and generalization capabilities, 10-fold cross-validation was performed on the training data using these optimal hyperparameters, yielding an average F1 score of 73.52%, as detailed in Table 12.

Table 12: The results of 10-fold cross-validation for the mention-pair model using the best hyper-parameters

precision	recall	F1
85.65	64.46	73.52

Hybrid Model: The same grid search procedure was employed for the hybrid model with the random forest classifier. The optimal hyperparameters for this model were found to be:

- **Max Depth:** 30
- **Number of Estimators:** 500
- **Criterion:** Entropy

Table 13 presents the results of 10-fold cross-validation on the training data, demonstrating the impact of incorporating word embeddings on model performance.

Table 13: 10-fold cross-validation results for the hybrid model using the best hyper-parameters

	Precision	Recall	F1
Feature vector results without considering word embedding	88.71	81.33	84.85
Feature vector results with considering word embedding	91.17	79.63	84.96

A two-hidden-layer architecture was employed for the hybrid model utilizing the MLP classifier, with the first layer consisting of 1500 nodes and the second layer containing 300 nodes. The ReLU activation function was used in both hidden layers. The optimal hyperparameters for this model, obtained through experimentation and fine-tuning, are summarized in Table 14. The training results for this configuration, as reported in Table 15 , demonstrate the effectiveness of the MLP classifier in conjunction with our hybrid approach.

Table 14: Optimal hyper-parameters of MLP hybrid model

First Hidden layer	
Kernel initializer	identity
Regularization value	0.001
dropout	0.6
Second Hidden layer	
Kernel initializer	identity
Regularization value	0.001
dropout	0.7
Other Model Settings	
Learning rates	0.0001
Optimization Algorithm	Adam

Table 15: 10-fold cross-validation results for the hybrid MLP model using the best parameters

Precision	Recall	F1
80.22	86.89	81.89

5-2- Experiments on RCDAT Corpus

This section evaluates our proposed hybrid model on the established RCDAT corpus, focusing on pronouns that are part of an existing coreference chain. Following the methodology of previous studies on this corpus, we utilize the Uppsala test documents for evaluation, comprising 415 sentences and 16,274 tokens. As with the Mehr corpus, we assess the model's performance using random forest and MLP classifiers across various configurations, with the results summarized in
Table 16. Section 5.2.1 provides an in-depth analysis of the RCDAT results, while Section 5.2.2 details the hyper-parameter tuning process. It is crucial to emphasize that the evaluation of the RCDAT corpus expressly excludes non-referential pronouns and pronouns initiating new chains, as these annotations are not available in the corpus. This focused approach allows for direct comparison with previous research on the RCDAT corpus and provides a clear assessment of the model's ability to resolve pronouns within established coreference chains.

Table 16: The reported results of the presented system on RCDAT corpus compared to previous works and various settings

	System	Using Rule-Based Modules	GloVe/BERT	Cluster-Based Features	Precision	Recall	F1
1	Rule-based baseline	+	-	-	42.66	41.15	41.89
2	Mention-pair baseline (random forest)	-	-	-	50.91	48.67	49.5
3	(random forest) [40]	-	GloVe	-	43.66	42.2	42.88
4	(MLP) [40]	-	GloVe	-	49.75	46.63	48.14
	Proposed System with Different Settings (Random Forest)						
5		-	-	system	41.83	40.35	41.08
6		+	-	system	49.8	48.2	49
7		+	GloVe	system	51.37	49.55	**50.45**
8		-	-	gold	53.76	51.85	52.79
9		+	-	gold	55.22	53.27	54.23
10		+	GloVe	gold	56.05	54.07	**55.04**
11		+	BERT	gold	58.48	56.09	**57.26**
12		+	BERT	system	53.15	51.19	**52.15**
	Proposed System with Different Settings (MLP)						
13		+	GloVe	gold	49.34	48.5	48.92
14		+	GloVe	system	46.78	45.13	45.94
15		+	BERT	system	48.90	46.48	47.66

5-2-1- RCDAT Corpus Result Analysis

The experimental results on the RCDAT corpus, as presented in
Table 16, offer valuable insights into the efficacy of our hybrid model and its various configurations in resolving pronouns within established coreference chains. It is important to note that the RCDAT corpus lacks annotations for non-referential pronouns and pronouns initiating new chains. Thus, our evaluation focuses on pronouns with existing antecedents within the corpus. This allows for a direct comparison with previous work on the RCDAT corpus and provides a clear assessment of the model's performance in resolving pronouns within existing discourse contexts.

Hybrid Model's Consistent Superiority: The proposed hybrid model outperforms both the rule-based baseline (Line 1) and the mention-pair baseline (Line 2) across all configurations. This underscores the effectiveness of combining rule-based knowledge with machine-learning techniques for Persian pronoun resolution, even on a different corpus (RCDAT), compared to the previous section's focus on Mehr.

Impact of BERT Embeddings (Gold Clusters): When using gold cluster features, the addition of BERT embeddings consistently leads to substantial improvements in F1 score (57.26 vs. 52.79, Line 11 vs. Line 8), highlighting the crucial role of contextualized representations in accurately resolving pronouns within established coreference chains.

Significance of GloVe Embeddings (System Clusters): In configurations utilizing system-generated cluster features, the inclusion of GloVe embeddings provides a notable boost to F1 score (50.45 vs. 49.00, Line 7 vs. Line

6), indicating that these embeddings capture valuable semantic and syntactic relationships that aid in pronoun resolution, even when gold-standard cluster information is not available.

Rule-Based Modules' Continued Relevance: While BERT embeddings offer significant advantages, the inclusion of rule-based modules still positively impacts performance, especially when using system-generated cluster features. This suggests that rule-based knowledge remains valuable for resolving certain types of pronoun references, even when advanced contextual embeddings are available.

Random Forest vs. MLP: Similar to the results on the Mehr corpus, the random forest classifier outperforms the MLP classifier across comparable settings. This reinforces that the random forest's ability to model complex feature interactions is particularly well-suited for pronoun resolution tasks.

Challenges of Automatic Feature Extraction: The performance gap between configurations using gold and system-generated cluster features underscores the continued need for improvement in automatic feature extraction techniques for Persian coreference resolution. This suggests that the quality of automatically generated features still lags behind gold standard annotations, impacting the performance of the hybrid model in real-world scenarios.

Potential for Further Improvement: While our hybrid model demonstrates significant advancements in Persian pronoun resolution, there remains room for further improvement. This could involve exploring more sophisticated feature engineering techniques, refining the rule-based modules, and investigating alternative machine learning architectures that can better leverage the information provided by contextualized embeddings like BERT.

5-2-2- Hyper-Parameter Settings for Implemented Systems on RCDAT Corpus

Hyper-parameter tuning is a critical step in optimizing the performance of machine learning models. In this section, we detail the hyper-parameter optimization process for our Persian pronoun resolution models on the RCDAT corpus. For all experiments, the antecedent candidate selection was restricted to a sentence window of 3, and the merging threshold was set to 0.2, as determined during the development phase.

Mention-Pair Model: The optimal hyperparameters for the random forest classifier in the mention-pair baseline were determined through a grid search, as outlined in Table 11. The best configuration, yielding an average F1 score of 69.00 in 10-fold cross-validation on the training data (Table 17), was:

- **Max Depth:** 40
- **Number of estimators:** 1000
- **Criterion:** Entropy

Table 17: The results of 10-fold cross-validation for the mention-pair model using the best hyper-parameters

Precision	Recall	F1
84.19	58.46	69

Hybrid Model: Similarly, a grid search was employed to optimize the hyperparameters of the random forest classifier in the hybrid model. The optimal configuration, resulting in an average F1 score of 84.96 in 10-fold cross-validation (Table 18), was:

- **Max Depth:** 50
- **Number of Estimators:** 1000
- **Criterion:** Entropy

Table 18: 10-fold cross-validation results for the hybrid model using the best hyper-parameters

	Precision	Recall	F1
Feature vector results without considering word embedding	88.75	81.48	84.95
Feature vector results with considering word embedding	88.79	81.46	84.96

The hybrid model with the MLP classifier utilized a three-layer architecture with 850, 680, and 440 nodes, respectively, using the ReLU activation function. Table 19 details the optimal hyperparameters for this model, including kernel initializers, regularization values, and dropout rates for each layer. The learning rate was set to 0.0001, and the Adam optimizer was employed for training.

Table 19: Optimal hyper-parameters of hybrid MLP model

First Hidden layer	
Kernel initializer	glorot_normal
Regularization value	0.001
dropout	0.5
Second Hidden layer	
Kernel initializer	identity
Regularization value	0.01
dropout	0.1

Third Hidden layer	
Kernel initializer	identity
Regularization value	0.001
dropout	0.1
Other Model Settings	
Learning rates	0.0001
Optimization Alghorithm	Adam

6- Discussion and Implications

6-1- Comparative Analysis and Key Findings

Our experimental results on the Mehr and RCDAT corpora (Tables 9 & 16) provide compelling evidence for the effectiveness of our hybrid model in Persian pronoun resolution. Our approach consistently outperforms both rule-based and mention-pair baselines, underscored by the value of integrating linguistic knowledge with machine learning, especially when enhanced with contextualized embeddings like BERT and GloVe.
The results further highlight several key findings:

The Impact of Contextual Embeddings:
The incorporation of BERT embeddings, in particular, significantly boosts performance, demonstrating the importance of capturing contextual nuances for accurate pronoun resolution in Persian. Using embeddings further enhances accuracy, suggesting a complementary relationship in capturing different linguistic information.

The Results Further Highlight Several Key Findings:
The Impact of Contextual Embeddings: The incorporation of BERT embeddings, in particular, significantly boosts performance, demonstrating the importance of capturing contextual nuances for accurate pronoun resolution in Persian. The combination of BERT with GloVe embeddings further enhances accuracy, suggesting a complementary relationship in capturing different linguistic information.

The Value of Rule-Based Modules: While BERT Embeddings play a crucial role, incorporating rule-based modules remains beneficial, especially for handling specific linguistic patterns and more straightforward coreference cases.

The Superiority of Random Forest: The random forest classifier consistently outperforms the MLP classifier, highlighting its effectiveness in modeling the complex interactions among diverse features.

6-2- Broader Implications and Contributions

The implications of our research extend beyond the Persian language. The hybrid approach presented in this paper offers a versatile framework for pronoun resolution that can be adapted to other under-resourced languages. By combining the strengths of rule-based and machine-learning methods and incorporating contextualized embeddings, our model provides a practical and interpretable solution for addressing the challenges of pronoun resolution in diverse linguistic contexts.
Furthermore, the development of the Mehr corpus represents a significant contribution to the Persian NLP community. This comprehensive resource, meticulously annotated with gold-standard coreference chains and enriched with automatically labeled linguistic elements, provides a valuable benchmark for future research and development in Persian coreference resolution.

6-3- Limitations and Future Directions

While our hybrid model demonstrates promising results, several limitations warrant further investigation. The performance gap between configurations using gold and system-generated cluster features highlights the need for continued research in improving automatic feature extraction methods. Additionally, exploring alternative machine learning architectures, such as transformer-based models specifically tailored for Persian pronoun resolution, may yield further performance gains.
In future work, we plan to investigate the effectiveness of our hybrid approach in other less-studied languages and explore its applicability to specific domains, such as legal and medical texts. We also aim to develop more sophisticated techniques for integrating external knowledge sources, such as knowledge graphs, to enhance the model's ability to resolve ambiguous pronoun references. Furthermore, we will explore methods to mitigate potential biases in the model's predictions, ensuring fairness and equity in its application.

7- Conclusion

This paper presents a novel hybrid entity-centric model for Persian pronoun resolution, effectively addressing the challenges posed by the need for more specialized resources and complex linguistic features in under-resourced languages. Our approach synergistically integrates rule-based modules, capturing linguistic patterns, with a machine-learning module enhanced by Persian BERT embeddings to leverage contextual information. This combination has resulted in significant

performance improvements, surpassing existing Persian systems on both the Mehr and RCDAT corpora, with a notable 4.68 F1-score increase on the Mehr corpus and a 4.01 increase on the RCDAT corpus.

The introduction of the Mehr corpus, a comprehensive Persian coreference dataset annotated with a wide range of linguistic phenomena, significantly contributes to the research community. This corpus addresses the limitations of previous resources and provides a valuable benchmark for future research in Persian coreference resolution.

The performance of our hybrid model, particularly when incorporating BERT embeddings and cluster-level features, demonstrates the efficacy of combining diverse linguistic cues and contextual information for accurate pronoun resolution. Furthermore, the model's modularity, interpretability, and computational efficiency make it a practical solution for real-world applications.

Future work will focus on refining automatic feature extraction methods, exploring alternative machine-learning architectures, and incorporating additional knowledge sources to enhance performance further. We also aim to investigate the generalizability of our approach to other under-resourced languages and to adapt it to specific domains, such as legal and medical texts. By addressing these challenges, we strive to advance the state of the art in pronoun resolution, particularly for languages with limited resources, and contribute to developing more comprehensive and accurate natural language processing systems.

References

[1] Lee, K., et al. End-to-end Neural Coreference Resolution. in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017. Association for Computational Linguistics.

[2] Ng, V. and C. Cardie. Improving machine learning approaches to coreference resolution. in Proceedings of the 40th annual meeting on association for computational linguistics. 2002. Association for Computational Linguistics.

[3] Bahdanau, D., K.H. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. in 3rd International Conference on Learning Representations, ICLR 2015. 2015.

[4] Kriman, S. and H. Ji. Joint detection and coreference resolution of entities and events with document-level context aggregation. in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop. 2021.

[5] Liu, Z., K. Shi, and N. Chen. Coreference-Aware Dialogue Summarization. in Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2021.

[6] Mao, R. and X. Li. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. in Proceedings of the AAAI conference on artificial intelligence. 2021.

[7] Vu, H.H., H. Kamigaito, and T. Watanabe, Context-Aware Machine Translation with Source Coreference Explanation. Transactions of the Association for Computational Linguistics, 2024. 12: p. 856-874.

[8] Bohnet, B., C. Alberti, and M. Collins, Coreference Resolution through a seq2seq Transition-Based System. Transactions of the Association for Computational Linguistics, 2023. 11: p. 212-226.

[9] Devlin, J., 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). 2019.

[10] Hürriyetoğlu, A., et al., Event Coreference Resolution for Contentious Politics Events. arXiv preprint arXiv:.10123, 2022.

[11] Lu, Y., et al., End-to-end neural event coreference resolution. Artificial Intelligence, 2022. 303: p. 103632.

[12] Yu, X., W. Yin, and D. Roth. Pairwise Representation Learning for Event Coreference. in Proceedings of the 11th Joint Conference on Lexical and Computational Semantics. 2022.

[13] Zhang, W., S. Wiseman, and K. Stratos. Seq2seq is All You Need for Coreference Resolution. in The 2023 Conference on Empirical Methods in Natural Language Processing.

[14] Wan, B., et al., A deep neural network model for coreference resolution in geological domain. Information Processing & Management, 2023. 60(3): p. 103268.

[15] Žabokrtský, Z., et al. Findings of the second shared task on multilingual coreference resolution. in Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution. 2023.

[16] Bi, M., et al., Machine reading comprehension combined with semantic dependency for Chinese zero pronoun resolution. Artificial Intelligence Review, 2023. 56(8): p. 7597-7612.

[17] Demir, Ş., Neural coreference resolution for Turkish. Journal of Intelligent Systems: Theory and Applications, 2023. 6(1): p. 85-95.

[18] Lu, J. and V. Ng. Learning antecedent structures for event coreference resolution. in 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). 2017. IEEE.

[19] Hobbs, J.R., Resolving pronoun references. Lingua, 1978. 44(4): p. 311-338.

[20] Carbonell, J.G. and R.D. Brown. Anaphora resolution: a multi-strategy approach. in Proceedings of the 12th conference on Computational linguistics-Volume 1. 1988. Association for Computational Linguistics.

[21] Lappin, S. and H. Leass, An algorithm for pronominal anaphora resolution. Computational linguistics, 1994. 20(4): p. 535-561.

[22] Kennedy, C. and B. Boguraev. Anaphora for everyone: pronominal anaphora resoluation without a parser. in Proceedings of the 16th conference on Computational linguistics-Volume 1. 1996. Association for Computational Linguistics.

[23] Stuckardt, R., Design and enhanced evaluation of a robust anaphor resolution algorithm. Computational Linguistics, 2001. 27(4): p. 479-506.

[24] Haghighi, A. and D. Klein. Simple coreference resolution with rich syntactic and semantic features. in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. 2009. Association for Computational Linguistics.
[25] Raghunathan, K., et al. A multi-pass sieve for coreference resolution. in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010. Association for Computational Linguistics.
[26] Lee, H., et al. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. in Proceedings of the fifteenth conference on computational natural language learning: Shared task. 2011. Association for Computational Linguistics.
[27] Lee, H., et al., Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics, 2013. 39(4): p. 885-916.
[28] Van Cranenburgh, A., et al. A Hybrid Rule-Based and Neural Coreference Resolution System with an Evaluation on Dutch Literature. in Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference. 2021.
[29] Lee, K., L. He, and L. Zettlemoyer. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018.
[30] Joshi, M., et al., Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 2020. 8: p. 64-77.
[31] Zhang, H., et al., Knowledge-aware Pronoun Coreference Resolution. arXiv preprint arXiv:.03663, 2019.
[32] Zhang, H., Y. Song, and Y. Song, Incorporating Context and External Knowledge for Pronoun Coreference Resolution. arXiv preprint arXiv:.10238, 2019.
[33] Hou, F., et al., Anisotropic span embeddings and the negative impact of higher-order inference for coreference resolution: An empirical analysis. Natural Language Engineering, 2024: p. 1-22.
[34] De Langhe, L., O. De Clercq, and V. Hoste, Constructing a cross-document event coreference corpus for Dutch. Language Resources and Evaluation, 2022: p. 1-30.
[35] Rahimi, Z. and S. HosseinNejad, Corpus based coreference resolution for Farsi text Signal and Data Processing2020. 17(1): p. 79-98.
[36] Haji Mohammadi, H., et al., Mehr: A Persian Coreference Resolution Corpus. Journal of AI and Data Mining, 2023.
[37] Biau, G. and E. Scornet, A random forest guided tour. Test, 2016. 25: p. 197-227.
[38] Artari, V.K.P., et al. A Multi-Pass Sieve Coreference Resolution for Indonesian. in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021.
[39] Sahlani, H., M. Hourali, and B. Minaei-Bidgoli, Coreference resolution with deep learning in the persian language. Signal and Data Processing2020.
[40] Sahlani, H., M. Hourali, and B. Minaei-Bidgoli, Coreference Resolution Using Semantic Features and Fully Connected Neural Network in the Persian Language. International Journal of Computational Intelligence Systems, 2020.
[41] Soon, W.M., H.T. Ng, and D.C.Y.J.C.l. Lim, A machine learning approach to coreference resolution of noun phrases. Computational linguistics, 2001. 27(4): p. 521-544.
[42] Ng, V. and C. Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. in COLING 2002: The 19th International Conference on Computational Linguistics. 2002.
[43] Denis, P. and J. Baldridge. Specialized models and ranking for coreference resolution. in Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008. Association for Computational Linguistics.
[44] Denis, P. and J. Baldridge, Global joint models for coreference resolution and named entity classification. Procesamiento del Lenguaje Natural, 2009. 42.
[45] Wiseman, S.J., et al. Learning anaphoricity and antecedent ranking features for coreference resolution. in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. Association for Computational Linguistics.
[46] Lee, K., et al., End-to-end neural coreference resolution. arXiv preprint arXiv:.05365, 2017.
[47] Luo, X., et al. A mention-synchronous coreference resolution algorithm based on the bell tree. in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. 2004. Association for Computational Linguistics.
[48] Nicolae, C. and G. Nicolae. Bestcut: A graph algorithm for coreference resolution. in Proceedings of the 2006 conference on empirical methods in natural language processing. 2006. Association for Computational Linguistics.
[49] Culotta, A., M. Wick, and A. McCallum. First-order probabilistic models for coreference resolution. in Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. 2007.
[50] Lee, H., M. Surdeanu, and D. Jurafsky, A scaffolding approach to coreference resolution integrating statistical and rule-based models. Natural Language Engineering, 2017. 23(5): p. 733-762.
[51] Lai, T.M., T. Bui, and D.S. Kim. End-to-end neural coreference resolution revisited: A simple yet effective baseline. in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. IEEE.
[52] Zhang, R., et al., Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. arXiv preprint arXiv:1805.04893, 2018.
[53] Joshi, M., et al. BERT for Coreference Resolution: Baselines and Analysis. in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.
[54] Kantor, B. and A. Globerson. Coreference resolution with entity equalization. in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
[55] Werlen, L.M. and J. Henderson. Graph Refinement for Coreference Resolution. in Findings of the Association for Computational Linguistics: ACL 2022. 2022.

[56] Zhang, W., S. Wiseman, and K. Stratos, Seq2seq is all you need for coreference resolution. arXiv preprint arXiv:2310.13774, 2023.
[57] Guo, Q., et al. Dual Cache for Long Document Neural Coreference Resolution. in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.
[58] Moosavi, N.S. and G. Ghassem-Sani, A ranking approach to Persian pronoun resolution. Advances in Computational Linguistics. Research in Computing Science, 2009. 41: p. 169-180.
[59] Denis, P. and J. Baldridge. A Ranking Approach to Pronoun Resolution. in IJCAI. 2007.
[60] Lee, K., L. He, and L. Zettlemoyer, Higher-order coreference resolution with coarse-to-fine inference. arXiv preprint arXiv:.05392, 2018.
[61] Chai, H. and M. Strube. Incorporating Centering Theory into Neural Coreference Resolution. in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022.
[62] Miculicich, L. and J. Henderson, Graph Refinement for Coreference Resolution. arXiv preprint arXiv:.16574, 2022.
[63] Wang, Y., Y. Shen, and H. Jin. An End-To-End Actor-Critic-Based Neural Coreference Resolution System. in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. IEEE.
[64] Hajimohammadi, H., A. Talebpour, and A. Mahmoudi Aznaveh. Using decision tree to Persian pronoun resolution. in Twenty-first annual National Conference of the Computer Society of Iran. 2016. Tehran: Computer Society of Iran.
[65] Nourbakhsh, A. and M. Bahrani. Persian Pronoun Resolution Using Data Driven Approaches. in International Conference on Information and Software Technologies. 2017. Springer.
[66] Fallahi, F. and M. Shamsfard, Recognizing anaphora reference in Persian sentences. International Journal of Computer Science Issues (IJCSI), 2011. 8(2): p. 324.
[67] Nazaridoust, M., B.M. Bidgoli, and S. Nazaridoust. Co-reference Resolution in Farsi Corpora. in In Advance Trends in Soft Computing: Proceedings of WCSC 2013. 2013. Cham: Springer International Publishing.
[68] Asgari-Bidhendi, M., A. Hadian, and B. Minaei-Bidgoli, Farsbase: The persian knowledge graph. Semantic Web, 2019. 10(6): p. 1169-1196.
[69] Bijankhan, M., The role of the corpus in writing a grammar: An introduction to a software. Iranian Journal of Linguistics, 2004. 19(2): p. 48-67.
[70] Mirzaei, A. and P. Safari. Persian Discourse Treebank and coreference corpus. in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.
[71] Yimam, S.M., et al. WebAnno: A flexible, web-based and visually supported system for distributed annotations. in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2013.
[72] Hirschman, L., MUC-7 coreference task definition, version 3.0. Proceedings of MUC-7,, 1997.
[73] Bergsma, S. and D. Lin. Bootstrapping path-based pronoun resolution. in Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics. 2006.
[74] Ji, H. and D. Lin. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. in Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1. 2009.
[75] Rasooli, M.S., M. Kouhestani, and A. Moloodi. Development of a Persian syntactic dependency treebank. in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.
[76] Seraji, M., B. Megyesi, and J. Nivre, Bootstrapping a Persian dependency treebank. Linguistic Issues in Language Technology, 2012. 7(18).

Journal of Information Systems and Telecommunication
Vol.14, No.4, July-September 2026, 268-276
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Hybrid Embedding and GPT-5 Instruction-Augmented Framework for Arabic Hate Speech Detection

Anas Khudhur Abbas al juboori[1], Heshaam Faili[2*]

[1]. Computer Engineering, Department Pardis Alborz, University of Tehran, Tehran, Iran
[2]. School of Electrical and Computer Engineering, College of Engineering, University of Tehran , Tehran, Iran

Received: 10 Oct 2025/ Revised: 04 Sep 2025/ Accepted: 07 Oct 2026

Abstract
Hate speech is defined as language that creates hatred, hostility, or violence towards people or groups based on their religion, ethnicity, gender, and/or nationality. As social media continues to grow and expand, the influx of hate speech and abusive language on these platforms is on the rise and has become a significant threat to online safety and social cohesion. Despite improvements in the detection of hate speech with automation technologies, there is little research conducted on the detection of hate speech in Arabic, especially regarding varying dialects. Current approaches often fail to identify language patterns or contextual nuances that are specific to different dialects. Moreover, little research has been conducted focused solely on instruction-based learning in this field. We presented a methodology for detecting abusive language and hate speech in Arabic utilizing the L-HSAB database. Through the combination of AraBERT-v2 and Multilingual MPNet, hybrid embeddings were created that have both language-specific characteristics and cross-linguistic conceptual characteristics. UMAP and HDBSCAN were applied for both dimension reduction and cluster analysis, and iterative refinement of clustering was facilitated by the utilization of Silhouette and Coherence scores. Based on the identified resulting clusters generated, GPT-5 was used to produce linguistically grounded instructional prompts, which were then assessed by both LLM-as-a-Judge using prompt generation and by human validators. Finally, the identified instructions were incorporated into further fine-tuning of the models. Experimental results indicate that the proposed method yields 90.5% accuracy for the purpose of detecting hate speech in dialects of Arabic.

Keywords: Hate Speech; Large Language Model; Supervised Fine Tuning; Arabic.

1- Introduction

The rapid growth of the internet and social media has reshaped global communication, with over 5.56 billion internet users and 5.24 billion social media users worldwide in 2025. These platforms have become central to digital discourse, but alongside positive engagement, they have also facilitated the spread of harmful content such as hate speech and abusive language, often targeting sensitive attributes like religion, ethnicity, or gender [1]. In response, regulatory initiatives such as the EU hate speech code seek to address the issue, yet automated detection remains difficult due to the massive scale of online content and limitations of existing tools [2]. Although Arabic ranks among the most widely used online languages with more than 362 million native speakers [3], research on hate speech detection has largely focused on high-resource languages like English and Spanish, and Arabic receives far less attention [4].

Arabic hate speech detection remains underexplored, especially in dialects like Levantine. Existing resources are scarce and mostly centered on Modern Standard Arabic (MSA), which does not capture the informal, dialectal language of social media. Moreover, traditional methods struggle with noisy, imbalanced data and overlook linguistic cues, creating a need for dialect-specific and context-aware approaches [5,6]. The challenges related to the detection of Arabic hate speech have motivated the methodological decisions made in this research study. The first challenge pertains to the vast diversity of dialects in the Arabic language. Moreover, social media users often switch between Arabic and another language. Problems with

✉ **Corresponding Author**
hfaili@ut.ac.ir

having only one representation model and this linguistic diversity exist in that such a model may never encapsulate all of the pertinent semantic meaning and context. As a result, hybrid embedding can be used to combine Arabic-only embeddings, as well as multilingual embeddings (i.e., English, French, Spanish, etc.), in order to address this problem. The second challenge is the amount of noise and heterogeneity present in social media data, resulting in a reduced ability of traditional, supervised methods to detect (i.e., accurately identify) hate speech. To address this issue, clustering techniques are employed to identify the underlying semantic structure of social media data and cluster similar instances together prior to formal training of the model. The third challenge is that instruction-based learning has been shown to improve both understanding and generalization of tasks using LLMs recently. However, instruction-based learning has not yet been adequately explored for the problem of hate speech detection in Arabic. Thus, a unified model that utilizes hybrid embeddings, clustering, and instruction-based fine-tuning should be developed to meet the requirements of dialectical Arabic. In this study, we aim to develop a robust methodology for hate speech detection, leveraging hybrid embeddings, adaptive clustering, and instruction-based fine-tuning to address the limitations of existing approaches. Our research contributions can be summarized as follows:

- We focus on Levantine Arabic, one of the least studied yet widely spoken Arabic dialects, by employing the L-HSAB dataset, a benchmark corpus specifically designed for abusive and hate speech detection.
- We propose a hybrid embedding strategy that combines AraBERT-v2 with Multilingual MPNet, enabling both language-specific and cross-lingual contextual understanding.
- We introduce an adaptive clustering (UMAP + HDBSCAN) with iterative validation (Silhouette and Coherence scores), and leverage clustering outputs for the generation of linguistically grounded instructions via GPT-5.
- We develop a dual-layer evaluation framework (LLM-as-a-Judge + human review) to validate and filter high-quality instructions.
- We demonstrate that instruction-augmented Supervised Fine-Tuning (SFT+INST) substantially enhances performance in detecting hate speech in Arabic.

The remainder of this paper is structured as follows: Section 2 reviews the related works. Section 3 details the data and methodology employed in our study. Section 4 provides an evaluation of our results. Finally, Section 5 highlights our work's key contributions and suggests future research directions.

2- Related Works

Several studies have addressed hate speech detection in Arabic, each focusing on different aspects of the problem. Doghmash and Saad [7] approached Arabic hate speech detection and masking on social media. They used the SemEval-2020 Arabic offensive language dataset (10,000 tweets) to compare deep learning architectures to transformer-based models for the task of hate speech detection. While traditional models achieved reasonable accuracy of ~80%, and simpler transformer models were shown to outperform them, the QARiB model utilizing pre-trained AraBERT was able to achieve 92% macro F1 and 95% accuracy to surpass all prior results. In addition to hate speech detection, the authors also proposed hate speech masking as a machine translation task, where they generated a parallel corpus that replaced the offensive term with a masked token. Their transformer-based model was evaluated with BLEU scores of 0.30 (1-gram).

Salloum et al. [8] studied Arabic hate speech in the context of Twitter using the L-HSAB dataset, focusing on the Levantine dialects. The methodology incorporated text preprocessing, TF-IDF representation, dimensionality reduction using PCA, and clustering with K-Means. This study grouped tweets into meaningful clusters, demonstrating intrinsic properties of hate speech within the data that might not be visible to unsupervised analysis alone.

Ahmed et al. [9] examined the barriers to detecting hate speech within Arabic, which encompasses the dialect continuum across Syria, Jordan, Palestine, and Lebanon, using the L-HSAB dataset. Their experiments comparing the various embedding methods showed that models trained on Arabic, like fastText and TF-IDF, performed considerably better than English embeddings, such as GloVe and GoogleNews Word2Vec, which performed near zero.

Charfi et al. [10] proposed the ADHAR dataset in response to the identified gaps in previous Arabic hate-speech resources, which included either a single dialect or a significant imbalance in the proportions of hate-speech. ADHAR contains a total of 4,240 tweets from five different dialects that were annotated for hate speech as binary categories and target sources (nationality, religion, ethnicity, race) with a high degree of annotator agreement. Their experimental results showed that traditional models performed reasonably well, but fine-tuned AraBERT performed best in the detection of hate (≈0.94 F1).

Daouadi et al. [11] found Arabic hate-speech detection on Twitter useful to describe the task as a five-class classification problem (non-hate, general hate, racial hate, religious hate, sexist hate), annotated tweets. Their dataset included 11,634 tweets that they had annotated. Their approach was to fine-tune AraBERT and MARBERT and then ensemble those models. They attempted to address the imbalance by semi-supervised augmentation, combining externally labeled religious hate tweets and applying pseudo-labeling on an additional corpus using their best-performing model. The results produced in the paper showed ensembling

provided better performance than individual models, with the permutation producing a weighted F1 of ~85.6, with significant improvement in performance in minority classes of religious and sexist hate.

Aldjanabi et al. [12] introduced a Multi-Task Learning (MTL) framework utilizing pre-trained Arabic language models to improve performance on tasks related to offensive and hate speech in Arabic. Unlike single-task models, their MTL approach learns from multiple datasets simultaneously, allowing the model to capture both global and dataset-specific representations. They utilized several benchmark datasets, including OSACT-OFF, OSACT-HS, L-HSAB, and T-HSAB, and modified AraBERT and MarBERT for this purpose. The results showed that their MTL framework outperformed traditional single-task models, particularly with MarBERT, which better addressed dialectal variations in Arabic.

Research on Arabic hate speech detection can be divided into three major areas. The first area presents transformer models and fine-tuning methods as a means of improving performance [7,11,12]. Recent results demonstrate that Arabic language models (e.g., AraBERT, MarBERT) deliver strong results when using supervised methods. However, while these methods offer an approach to leverage the semantic structure of language to enhance model performance, they do not utilize the latent semantic structure of the data for their classification. The second area focuses on feature extraction and embedding approaches [9], demonstrating the importance of utilizing representations specific to the Arabic language to capture language characteristics. These investigations emphasize comparing different embedding methods but do not consider using complementary representations. The third area includes research focused on building datasets and conducting exploratory analysis of dialectical hate speech [8,10]. This area of research provides information and resources concerning Arabic dialects but utilizes traditional clustering techniques and representations.

While progress continues in each of these areas, there are still a number of gaps in the available literature. First, there have not yet been sufficient attempts made to develop both Arabic and multilingual equivalent representations to capture both dialect-specific and contextual-based semantic representations. Clustering methods have primarily been used for exploratory analysis and are not used as a guide for adapting downstream models. Finally, instruction-based learning, which has shown promise across various NLP tasks, remains a largely unexploited approach to detecting hate speech in Arabic. This study seeks to address these gaps by integrating hybrid embeddings, adaptive clustering, and instruction generation to form an integrated model for detecting hate speech in Levantine Arabic.

3- Methodology

We build our work on the integration of hybrid embeddings, adaptive outlier detection, and clustering with iterative validation to obtain multilingual instruction prompts that combine unsupervised clustering with SFT. The steps of our proposed approach are depicted in Figure 1. In this study, we performed a multiclass classification task utilizing the L-HSAB dataset, with the objective of categorizing tweets into three distinct classes: normal, hate, or abusive. As a first task, we created hybrid representations of text by combining AraBERT-v2 [13] with a multilingual MPNet model [14], taking advantage of both a language-specific and cross-lingual contextual features. We then utilized UMAP [15] to reduce the dimensionality of embeddings before clustering with HDBSCAN [16], a hierarchical density-based clustering algorithm. To validate the clustering process, we used silhouette and coherence scores to improve the reliability of the clusters as we tuned parameters until we achieved semantically similar clusters. From our identified clusters, we created sample instructions that provided linguistic to the Large Language Model (LLM) that distinguished between the target classes in a range of ways. Our created prompts were critically evaluated using automated LLM-based judgement and human judgement, and we retained the best prompts. We utilized the provided dataset and valid instructions to fine-tune the classification model, and we evaluated its performance using standard evaluation metrics of accuracy, precision, recall, and F1-score, improving upon baseline methods.

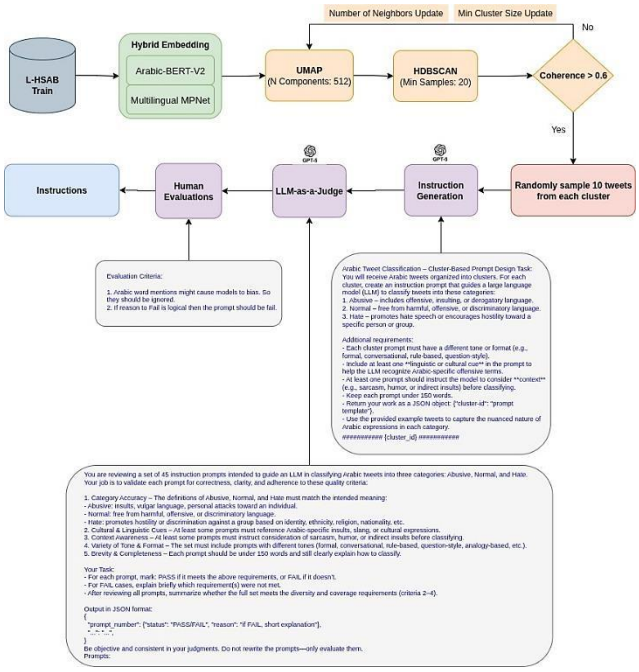


Fig. 1. The steps of our proposed approach.

3-1- Dataset

We utilized the L-HSAB dataset (Levantine Hate Speech and Abusive Language) [17], which is one of the first publicly available datasets aimed at hate speech detection for Levantine Arabic. L-HSAB is composed of tweets, gathered from Twitter and annotated into three categories: Normal, Abusive, and Hate. Table 1 shows the L-HSAB dataset statistics. L-HSAB is split into a training set and a testing set; the training set contains 1,369 abusive tweets, 2,968 normal tweets, and 339 hate tweets, which gives a total of 4,676 samples. In the testing set, there are 359 abusive, 682 normal, and 129 hate tweets, which is a total of 1,170 samples. As a whole, L-HSAB is an excellent starting point for research on hate speech and abusive language in Arabic. The strength of the dataset is that it specifically focuses on the Levantine dialect, which is less represented in existing multilingual corpora.

Table 1: Dividing the L-HSAB dataset.

Split	Abusive	Normal	Hate	Total
Train	1,369	2,968	339	4,676
Test	359	682	129	1,170

3-2- Hybrid Embedding Construction

To generate semantically enriched representations of the tweets, we employed a hybrid embedding strategy that integrates two complementary transformer-based models. The first model, AraBERT-v2, is specifically pre-trained on large-scale Arabic corpora and is optimized to capture morphological and syntactic properties unique to Arabic text. The second model, Multilingual MPNet-v2, is a multilingual sentence transformer designed to encode cross-lingual semantic similarity across diverse languages.

We chose both AraBERT-v2 and Multilingual MPNet due to their complementary abilities to represent dialectal Arabic content. AraBERT-v2 has been pretrained on a large scale using Arabic corpora, achieving exceptional results on multiple Arabic NLP tasks, whereas Multilingual MPNet has been designed specifically for language-agnostic semantic representations of multiple languages and can generalize contextually better than AraBERT-v2. Therefore, when using both models together, definitions of the Arabic language will include unique linguistic features as well as semantics that will allow for broader semantic relationships than what AraBERT would offer alone. We concatenated the embeddings from each model so that we could preserve all of the information generated from both models. If we were to average the embeddings of both models, we would force each model's embeddings into one common space and potentially diminish some of the features that pertain only to the dialect; thus, by concatenating the embeddings, we have preserved the unique features that were learned from each encoder. The end result of combining the two models created a representation with 1,536 features; however, the dimensionality reduction phase that followed reduced the computational cost while maintaining the most semantically informative structure for cluster analysis.

For each tweet, embeddings were independently computed using both models. These vectors were then horizontally concatenated (stacked) to form a unified feature vector that retains the strengths of both representations. AraBERT-v2 ensures coverage of Arabic-specific linguistic nuances, while Multilingual MPNet enriches contextual depth by embedding the tweet within a multilingual semantic space. This contextual enhancement strategy was designed to overcome the limitations of relying on a single embedding model, enabling the system to capture fine-grained semantic variations in Arabic tweets. The resulting hybrid embeddings served as the foundation for downstream clustering, validation, and classification tasks.

This process ultimately resulted in the creation of a large set of candidate instructions, from which only the most semantically coherent and linguistically rich prompts were retained for evaluation in the subsequent stage.

We used feature concatenation because it does not add extra fusion parameters and allows us to preserve all the information obtained from each encoder. Weighted fusion, on the other hand, needs pre-defined or learned weights, while concatenation makes use of the unique characteristics produced by both AraBERT-v2 and Multilingual MPNet at its original level. This feature allows the model to make use of the linguistic patterns in the Arabic language and enjoy some multilingual features as well. Although the use of concatenation creates a 1,536-dimensional feature vector, the usage of UMAP leads to a dimensionality reduction method that is efficient enough to compress the vector, taking into account its essential semantic structure for clustering.

3-3- Dimensionality Reduction and Clustering Validation

Given the high dimensionality of the hybrid embeddings, UMAP was first applied for dimensionality reduction. UMAP preserves local and global semantic structures while transforming the high-dimensional embeddings into a lower-dimensional space, thereby improving clustering efficiency. The complete parameter configurations used in UMAP and HDBSCAN, including their initial values, adjustment ranges, and final optimal values, are summarized in Table 2. The initial UMAP configuration included *n_neighbors=10*, *n_components=512*, and cosine similarity as the distance metric.

The rationale for using n_components = 512 was to maintain meaningful semantic information from 1,536-dimensional hybrid embeddings while reducing dimensionality for

improved clustering performance. Preliminary testing has demonstrated that larger amounts of dimensionality reduction have resulted in less coherent clusters. During refinement only n_neighbors and min_cluster_size were modified, as those two parameters have the most significant impact on building local neighborhoods and the degree of cluster granularity. All other parameters were held constant with the intention of decreasing the complexity of searches and maintaining consistency across iterations. A coherence threshold of 0.6 is to be utilized as an acceptable measure of semantic interpretability, meaning enough difference exists within the cluster content before being used to create instructional materials.

Following dimensionality reduction, we applied HDBSCAN for unsupervised clustering. HDBSCAN was chosen over K-Means due to its ability to automatically determine the number of clusters and to effectively handle heterogeneous datasets with varying density. Importantly, HDBSCAN assigns noisy or ambiguous data points as outliers (*label = -1*), ensuring that tweets without a clear semantic grouping are isolated rather than forcefully assigned to inappropriate clusters. The initial parameters for HDBSCAN included *min_cluster_size = 20*, *min_samples = 20*, *metric = euclidean*, and *cluster_selection_method = eom*.

To ensure semantic reliability of the clusters, we adopted a dynamic validation and refinement procedure. After each clustering run, two metrics were computed:

I) The Silhouette Score, which measures intra-cluster cohesion and inter-cluster separation, and
II) The Coherence Score, which evaluates the semantic interpretability of cluster contents.

If the coherence score of any cluster fell below 0.6, the clustering process was automatically re-executed with adjusted parameters. Specifically, *min_cluster_size* in HDBSCAN was randomly varied between *20* and *100*, while *n_neighbors* in UMAP was randomly varied between *10* and *200*. This process ensured that the clustering was iteratively refined until coherence thresholds were met.

The iterative optimization process stopped when all clusters reached the necessary coherence level of 0.6 or, in the case there were no changes after each algorithm optimization. During the optimization process, the parameters that were changed were only n_neighbors and min_cluster_size, as they were the only parameters that significantly changed the neighborhood maintenance process and contributed to the granularity level of the cluster, while other parameters were fixed to ensure the controlled and reproducible nature of the optimization process.

During experimentation, the initial run produced a Silhouette Score of *-0.156* and a Coherence Score of 0.452, indicating low semantic quality. After *52* iterations of parameter adjustment and re-clustering, the process converged to a solution with a Silhouette Score of *-0.018* and a Coherence Score of *0.636*, exceeding the required coherence threshold of *0.6*.

The final clustering results consisted of six coherent clusters alongside a substantial number of outliers (2,617 tweets), which were assigned to a separate group. Among the clusters, Cluster 0 was the largest, containing 1,049 tweets. These clusters, along with the outlier group, were subsequently used to guide instruction generation in later stages.

Table 2: Parameter settings for UMAP and HDBSCAN during clustering and refinement

Method	Parameter	Initial Value	Adjustment Range	Final Optimal Value
UMAP	n_neighbors	10	10 – 200	116
	n_components	512	–	512
	Metric	cosine	–	cosine
HDBSCAN	min_cluster_size	20	20 – 100	84
	min_samples	20	–	20
	Metric	euclidean	–	euclidean
	cluster_selection_method	eom	–	eom

Figure 2 illustrates the variation of Silhouette and Coherence scores across the 52 clustering iterations. Although it is true that the final Silhouette Score was nearly zero, this does not mean that the quality of clustering was poor. The characteristics of the sample datasets resulted in significant heterogeneity in the form of social media content, where normal, abusive, and hate speech have overlapping language constructs. In these types of environments, density-based clustering techniques also tend to produce lower silhouette scores due to there being no clear semantic distinction across different category types. Therefore, coherence scores were used as the primary metric for optimizing clustering, as it provides a measure of how semantically consistent the content found in each cluster is with each other. It should be noted that there wasn't a pre-established number of refinement iterations (52 iterations); this number was derived through the iterative optimization process through which each run satisfied the predetermined coherence level for the dataset in question. While the Silhouette score consistently remained around zero due to the heterogeneous and noisy nature of the dataset, the Coherence score showed significant improvement after repeated parameter adjustments, eventually exceeding the 0.6 threshold. This progression demonstrates the effectiveness of the iterative refinement strategy in achieving semantically meaningful clusters despite the challenging characteristics of the data.

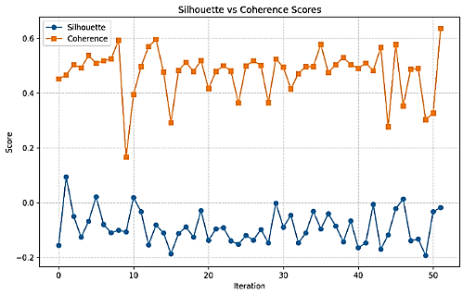


Fig. 2 Silhouette and Coherence scores across 52 clustering iterations.

3-4- Instruction Generation

Following the clustering and refinement stage, the next task was to create a series of instruction prompts for the classification model. To achieve this, a random sampling selection approach was used so that we randomly selected ten representative random samples of tweets belonging to each cluster. This ensured that the generated instructions captured the semantic diversity of each group while avoiding overfitting to a narrow set of examples. Outliers identified by HDBSCAN references were also grouped and considered in the instruction design as border ambiguous cases.

With the instructions generated from the sampled clusters, the next step was to use GPT-5 to assist with the instruction generation. GPT-5 was given the random sampling of clusters and was instructed to provide classification instructions designed to allow detection of Normal, Abusive, and Hate. Instructions were varied so that there was exploration of the various styles and directions the model was trained to guard against acceptance or overfitting to a single instruction type. The generated instructions included formal descriptions, conversational phrases, rule-based guidelines, and question-based directives, ensuring a broad coverage of styles.

Each instruction contained at least one linguistic element specific to Arabic, such as common abusive terms, culturally sensitive expressions, or region-specific slang. This was critical to expand the model's contextual understanding and limit the marginal ability to recognize hate speech or abusive speech spoken in Arabic. To maintain clarity and conciseness, all instructions were no more than 150 words, striking a balance between informativeness and brevity. This process ultimately resulted in the creation of a large set of candidate instructions, from which only the most linguistically rich and semantically coherent prompts were preserved for evaluation in the next phase.

3-5- Instruction Evaluation

The initial pool of instructions generated by GPT-5 was subjected to a rigorous multi-stage evaluation pipeline to ensure their reliability and effectiveness. The first phase was the LLM-as-Judge phase, where an LLM automatically assessed each instruction based on several predefined dimensions:

I) Category Accuracy or the degree to which the instruction correctly guides the model to differentiate between normal, abusive, and hate speech.

II) Linguistic cues, which measure the extent to which the instruction integrates relevant Arabic-specific abusive or hate-related expressions.

III) Contextual Awareness, or the degree to which the instruction addresses the nuance of meaning and borderline cases.

IV) Variation of Tone and Format, or variety of style across instructions; and,

V) brevity and completeness, confirming that the prompts remained under 150 words while being sufficiently informative.

Hence, automatically flagged instructions underperforming for one or more of these dimensions were discarded.

In the second stage, human evaluation was conducted to further refine the set of candidate instructions. Human annotators cross-verified the automated results and particularly focused on identifying instructions that contained explicit Arabic terms or culturally sensitive expressions that could lead to hallucinations or misinterpretations when used by the model. This step was crucial to prevent over-reliance on potentially misleading lexical markers.

To ensure that the evaluation process is as reliable as possible, a three-member team of independent annotators (all experienced in Arabic including linguistic analysis) conducted the human review stage. Each annotator provided their own review of the created instructions based on three major criteria (linguistic relevance, appropriateness to context, and conformity to the intended target category). Those that were either ambiguous, misleading, or incompatible with the intended category were discarded. Very little difference of opinion among the annotators occurred; the level of inter-annotator agreement was calculated using Fleiss' kappa coefficient ($\kappa = 0.78$), which demonstrates a substantial level of agreement between the evaluators. In cases where there were differences of opinion, the final decision on which instruction(s) to discard was made through discussion and majority agreement.

Through this two-tiered evaluation, the pool of candidate instructions was narrowed down to 21 high-quality prompts. To balance the dataset across categories, a subset of these instructions was selected: two were randomly assigned to the abusive category, while seven were retained for the hate category, ensuring that underrepresented classes received greater instructional coverage. The final selection was guided by performance metrics observed in preliminary zero-shot evaluations, where instructions were ranked according to their F1-score per class shown in Table 3. For example, instruction IDs [17, 19, 30, 35, 37, 42, 44] were chosen for the hate class based on their superior F1

performance, while instruction ID 36 was retained as the best performer for the abusive class. This procedure ensured that the final instruction set was not only diverse but also empirically validated for its effectiveness in guiding the classification model.

For the sake of hallucinations and biased output reduction, GPT-5 was making instructions only on the basis of tweets collected in each cluster and without using any other information. The instructions were evaluated using the framework of LLM-as-a-Judge according to their consistency regarding the category, relevance regarding the language, accuracy, and completeness. Finally, three separate human annotators checked the instructions and eliminated biased and misleading information.

Table 3: Ranking of instructions based on their F1-score in each class

	IDx	Abusive (F1-score)	Hate (F1-score)	Normal (F1-score)	Accuracy	Precision	Recall	F1-score
0	6	0.284069	0.29199	0.583732	0.420513	0.490735	0.50977	0.386597
1	8	0.485066	0.282828	0.62692	0.517949	0.483632	0.502789	0.464938
2	10	0.13964	0.278571	0.585227	0.390598	0.451822	0.482136	0.334479
3	14	0.49172	0.291262	0.682415	0.549573	0.503684	0.524856	0.488466
4	17	0.365482	0.296073	0.618215	0.463248	0.493004	0.517731	0.42659
5	19	0.437318	0.302285	0.613825	0.486325	0.493491	0.524255	0.451142
6	21	0.169565	0.289373	0.604186	0.413675	0.451872	0.490469	0.354374
7	23	0.458947	0.267606	0.518672	0.448718	0.482437	0.471891	0.415075
8	25	0.426554	0.290456	0.636522	0.501709	0.471005	0.499968	0.451177
9	30	0.514076	0.309859	0.671233	0.561538	0.502821	0.528698	0.498389
10	31	0.270073	0.275253	0.53	0.382906	0.463091	0.479884	0.358442
11	33	0.234615	0.275401	0.595149	0.412821	0.454409	0.478703	0.368389
12	34	0.51234	0.248996	0.445415	0.45812	0.499667	0.45929	0.40225
13	35	0.462703	0.314869	0.593284	0.500855	0.481938	0.49366	0.456952
14	36	0.528195	0.193548	0.614679	0.54188	0.511816	0.471156	0.445474
15	37	0.366492	0.322004	0.692053	0.523932	0.498211	0.534352	0.460183
16	39	0.366883	0.278003	0.520855	0.403419	0.486163	0.496193	0.38858
17	40	0.527596	0.19802	0.524793	0.483761	0.485903	0.463502	0.416803
18	42	0.405998	0.313646	0.509165	0.429915	0.464166	0.484573	0.409603
19	43	0.474895	0.253165	0.602996	0.503419	0.476112	0.47151	0.443685
20	44	0.46472	0.302521	0.652885	0.533333	0.4802	0.502119	0.473376

3-6- Model Training and Evaluation

LLAMA-3-8B was chosen for the final phase of SFT because it has an strong ability to follow instructions and competitive performance across in multilingual natural language understanding. The curated set of instructions amassed earlier was coupled with the labeled L-HSAB dataset during instruction-augmented supervised fine-tuning (SFT+INST) procedure. Through this coupling of the annotated examples with linguistically based instructions, the model learned to make task-specific distinctions while creating richer contextual representations.

4- Results and Discussion

The proposed approach was evaluated on the L-HSAB dataset using the SFT + INST. As shown in Table 4, the model achieved an accuracy of 90.5%, precision of 91.8%, recall of 85.3%, and an F1-score of 87.9%. These results confirm that

the combination of hybrid embeddings, adaptive clustering, and instruction-based guidance produces a competitive model.

For this paper, the results we report were derived from the official L-HSAB train-test data splits. Caution should be exercised when making direct comparisons between this work and previous studies since each may have used different data splits, preprocessing techniques, model configurations, and evaluation protocols. Therefore, the comparison shown in Table 4 was developed for general performance comparison purposes (not to be used as a strictly controlled benchmark).

We compared our proposed implementation with the study by Al-Jarrah et al. [18]. They explored an ensemble CNN-Bi-LSTM architecture and several transformer-based models. Their best results in the multi-class setting were obtained using Multilingual BERT, which achieved an F1-score of 80.0%, precision of 81%, recall of 79%, and accuracy of 81%. In contrast, our method substantially outperforms these results, reflecting a more balanced trade-off between precision and recall.

Additionally, Aldjanabi et al. [12] employed a cross-corpora multi-task learning framework incorporating datasets such as OSACT and L-HSAB, finding that the MarBERT-based multi-task model achieved an accuracy of 87.46% and an F1-score of 87.18% on these tasks. Although their multi-task setup improved over single-task methods, our single-task instruction-augmented model still edges ahead in both accuracy and F1-score, indicating the effectiveness of instruction-guided fine-tuning in dialectal contexts.

The performance enhancement noted is not only illustrated by the evaluation metrics but also by the ability of the presented model to deal with the major language hurdles that Arabic dialects face. The problems of key lexicon, spelling variation, code switching, and a low number of annotated texts are caused by the informal nature of this Arabic dialect. The hybrid embedding approach merges the benefits of AraBERT-v2 in terms of the representation of Arabic dialects, as it has an express understanding of the features related to the Arabic language, with multilingual semantic representations applied to the training of the Multilingual MPNet. The adaptive clustering organizes semantically similar tweets into the same group, which minimizes the negative influence of heterogeneous data. Hence, GPT-5 has more coherent information about the context. Therefore, this analysis shows how the fine-tuning method based on instructions allows the model to learn not only about the lexical features but also about the understanding of context and semantics, which are required for identifying hate speech in the dialectal varieties of Arabic.

Table 4: Evaluation and comparison of the results.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Al-Jarrah et al. [18] (Multilingual BERT)	81	81	79	80.0
Aldjanab et al. [12] (MarBERT)	87.46	-	-	87.18
Our Implementation (SFT + INST)	**90.5**	**91.8**	**85.3**	**87.9**

5- Conclusion

The new method developed in this research for identifying hate speech in Arabic employs the L-HSAB dataset. The new system uses AraBERT-v2 and a multilingual MPNet hybrid embedding, followed by adaptive clustering using UMAP and HDBSCAN. Clusters produced as a result were employed in the creation of instructions as part of SFT+INST. These instructions were created using the GPT-5 model and later verified using LLM-as-Judge and human reviewers. The resulting SFT+INST model achieved an accuracy of 90.5%, proving that merged hybrid semantic representations and instruction-based mechanisms used proved to be efficient in identifying hate speech in the Levantine Arabic dialect.

Despite achieving impressive results, several shortcomings must be highlighted. Firstly, instruction generation relies heavily on GPT-5, which might lead to model-specific biases. As a result, the computation cost of this system considerably rises, reducing accessibility in resource-constrained environments. Additionally, the clustering step produced several outliers, indicating that a part of the dataset could not be organized into a coherent group semantically. Thus, the performance of the system evidently heavily depends on the clustering stage. Another point to mention is that the suggested method has been tested using only one dataset of Levantine Arabic L-HSAB. Therefore, it is still unclear how effective this method might be in the case of application to other dialects of Arabic, multilingual cases, or different moderation instances.

Future research will focus on the context of the suggested model in relation to other Arabic dialects, and larger datasets will be examined as well. It would also be compulsory to replace the GPT-5 with Smaller Language Models (SLMs), which would come at a lower computational cost and ensure similar performance. Extending the mechanism to more content moderation aspects, like offensive and abusive language, etc., is also a promising perspective for research.

References

[1] K. Gasmi et al., "Hybrid Feature and Optimized Deep Learning Model Fusion for Detecting Hateful Arabic Content," IEEE Access, 2025.

[2] K. E. Daouadi, Y. Boualleg, and O. Guehairia, "Systematic investigation of recent pre-trained language model for hate speech

detection in arabic tweets," ACM Transactions on Asian and Low-Resource Language Information Processing, 2024.
[3] B. Corporation, "Most Spoken Languages in the World," 2025. [Online]. Available: https://www.berlitz.com/blog/most-spoken-languages-world.
[4] F. Husain and O. Uzuner, "A survey of offensive language detection for the arabic language," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 20, no. 1, pp. 1-44, 2021.
[5] M. M. Abdelsamie, S. S. Azab, and H. A. Hefny, "The dialects gap: A multi-task learning approach for enhancing hate speech detection in Arabic dialects," Expert Systems with Applications, vol. 295, p. 128584, 2026.
[6] A. Dahou et al., "A Survey on Dialect Arabic Processing and Analysis: Recent Advances and Future Trends," ACM Transactions on Asian and Low-Resource Language Information Processing, 2025.
[7] S. T. Doghmash and M. Saad, "Arabic Hate Speech Identification and Masking in Social Media using Deep Learning Models and Pre-trained Models Fine-tuning," arXiv preprint arXiv:2507.23661, 2025.
[8] S. Salloum, K. Tahat, A. Mansoori, R. Alfaisal, and D. Tahat, "Leveraging K-Means Clustering for Analysis of Arabic Hate Speech Tweets," in Global Congress on Emerging Technologies (GCET-2024), 2024: IEEE, pp. 282-285.
[9] A. H. Ahmed, R.-J. Yew, X. Minocher, and S. Venkatasubramanian, "Navigating Dialectal Bias and Ethical Complexities in Levantine Arabic Hate Speech Detection," arXiv preprint arXiv:2412.10991, 2024.
[10] A. Charfi, M. Besghaier, R. Akasheh, A. Atalla, and W. Zaghouani, "Hate speech detection with ADHAR: a multi-dialectal hate speech corpus in Arabic," (in English), Frontiers in Artificial Intelligence, Review vol. Volume 7 - 2024, 2024-May-30 2024, doi: 10.3389/frai.2024.1391472.
[11] K. E. Daouadi, Y. Boualleg, and K. E. Haouaouchi, "Ensemble of pre-trained language models and data augmentation for hate speech detection from Arabic tweets," arXiv preprint arXiv:2407.02448, 2024.
[12] W. Aldjanabi, A. Dahou, M. A. Al-Qaness, M. A. Elaziz, A. M. Helmi, and R. Damaševičius, "Arabic offensive and hate speech detection using a cross-corpora multi-task learning model," in Informatics, 2021, vol. 8, no. 4: MDPI, p. 69.
[13] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," arXiv preprint arXiv:2003.00104, 2020.
[14] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," arXiv preprint arXiv:1908.10084, 2019.
[15] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," arXiv preprint arXiv:1802.03426, 2018.
[16] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," J. Open Source Softw., vol. 2, no. 11, p. 205, 2017.
[17] H. Mulki, H. Haddad, C. B. Ali, and H. Alshabani, "L-hsab: A levantine twitter dataset for hate speech and abusive language," in Proceedings of the third workshop on abusive language online, 2019, pp. 111-118.
[18] H. Al-Jarrah, M. Al-Smadi, M. Hammad, and F. Shannaq, "Using Deep Learning Techniques to Detect Hate and Abusive Language in Arabic Tweets," International Journal of Intelligent Engineering & Systems, vol. 17, no. 5, 2024.