



PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
MOHAMED BOUDIAF UNIVERSITY - M'SILA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE



International Symposium on
Informatics and its Applications



ISIA 2024

December 10-11, 2024

The Sixth International Symposium on Informatics and its Applications

Proceedings Book





PROCEEDINGS BOOK OF THE SIXTH INTERNATIONAL SYMPOSIUM ON INFORMATICS AND ITS APPLICATIONS ISIA 2024

PROCEEDINGS BOOK OF ISIA 2024

December 10-11, 2024

Mohamed Boudiaf University of M'Sila, Algeria

Faculty of Mathematics and Computer Science



Preface

We are delighted to present the Proceedings of the Sixth International Symposium on Informatics and its Applications (ISIA 2024), held on December 10–11, 2024, at Mohamed Boudiaf University of M'Sila, Algeria. This prestigious symposium has once again provided an exceptional platform for researchers, academics, and practitioners to exchange groundbreaking findings, innovative ideas, and invaluable experiences in the ever-evolving field of computer science.

We extend our deepest gratitude to the contributing authors, whose valuable contributions enriched the symposium, to the diligent reviewers who meticulously assessed each submission, and to the dedicated organizing committee, whose unwavering efforts ensured the success of this event. We also extend our sincere appreciation to everyone who supported us throughout the various stages of this symposium.

We hope that this edition of the symposium will serve as an enriching opportunity for researchers and participants to engage deeply with the latest scientific advancements and to foster meaningful connections within the scientific community.

Sincerely,
ISIA'2024 Chairs

COMMITTEES

Honorary General Chair

Pr. BOUDELAA Amar, Rector of Mohamed Boudiaf University of M'sila

Vice Honorary General Chair

Pr. NOUIRI Brahimi, Dean of the Mathematics and Computer Science Faculty, Mohamed Boudiaf University of M'sila

General Chair

Dr. Mahmoud BRAHIMI

General Co-Chairs

Dr. Tahar MEHENNI

Pr. Hicham DEBBI

Scientific Committee Chair

Pr. Samir AKROUF

ORGANIZATION COMMITTEE

• Abdelbasset BARKAT	University of M'sila	• Mahmoud BRAHIMI	University of M'sila
• Adel MOUSSAOUI	University of M'sila	• Nasereddine AMROUNE	University of M'sila
• Ali DABBA	University of M'sila	• Nouredine AMRAOUI	University of M'sila
• Allaoua HEMMAK	University of M'sila	• Nouredine CHIKOUCHE	University of M'sila
• Belkacem BRAHIMI	University of M'sila	• Rabah MOKHTARI	University of M'sila
• Bilal LOUNNAS	University of M'sila	• Rached YAGOUBI	University of M'sila
• Hicham DEBBI	University of M'sila	• Roussafi MAHDJOUBI	University of M'sila
• Linda BELABDELOUAHAB	University of M'sila	• Said KADRI	University of M'sila
• Makhoul BENAZI	University of M'sila	• Salah GUESMIA	University of M'sila
• Mohamed BOUNIF	University of M'sila	• Samir AKROUF	University of M'sila
• Mohamed CHATRA	University of M'sila	• Samir ATHMANIA	University of M'sila
• Mohamed KAMEL	University of M'sila	• Tahar MEHENNI	University of M'sila

SCIENTIFIC COMMITTEE

- Pr. Abdelhak Bentaleb (Computer Science and Software Engineering, Concordia University, Canada)
- Pr. Abdelouahab Moussaoui (U. Setif 1, Algeria)
- Pr. Allel Hadjali (ENSMA, Poitiers, Nouvelle-Aquitaine, France)
- Pr. Benmohammed Mohamed (U. Abdelhamid Mehri Constantine 2, Algeria)
- Pr. Bilami Azzedine (U. Batna 2, Algérie)
- Pr. Mustapha Bourahla (U. M'sila, Algeria)
- Pr. Bouziane Abderraouf (U. Mohamed El Bachir El Ibrahim, BBA, Algérie)
- Pr. Chahir Youssef (University of Caen, France)
- Pr. Nouredine Chikouche (U. M'sila, Algeria)
- Pr. Djaafar Zouache (U. Mohamed El Bachir El Ibrahim, BBA, Algérie)
- Pr. Farid Meziane (University of Derby, UK)
- Pr. Farid Nouioua (U. Mohamed El Bachir El Ibrahim, BBA, Algérie)
- Pr. Mohamed Chaouki Babahenini (U. Biskra, Algeria)
- Pr. Mohamed Deriche (Ajman University, UAE)
- Pr. Morad Benyoucef (Telfer School of Management University of Ottawa, Canada)
- Pr. Mourad Amad (U. Bouira, Algeria)
- Pr. Salim Bouamama (U. Setif 1, Algeria)
- Pr. Chaker Abdelaziz Kerrache (U. Laghouat, Algeria)
- Pr. Saad Harous (College of Computing and Informatics, University of Sharjah, UAE)
- Pr. Samir A. El-Seoud (The British University in Egypt)
- Pr. Zahid Akhtar (State University of New York Polytechnic Institute, USA)
- Pr. Zitouni Abdelhafid (U. Abdelhamid Mehri Constantine 2, Algeria)
- Pr. Zsóka Zoltán (Budapest University of Technology and Economics, Hungary)
- Pr. Allaoua Hemmak (U. M'sila, Algeria)
- Pr. Hicham Debbi (U. M'sila, Algeria)
- Pr. Lamri Sayad (U. M'sila, Algeria)
- Pr. Abdellah Kouzou (U. Djelfa, Algeria)
- Dr. Abdelbasset Barkat (U. M'sila, Algeria)
- Dr. Abdelouahab Attia (U. BBA, Algérie)
- Dr. Ali Dabba (U. M'sila, Algeria)
- Dr. Adel Moussaoui (U. M'sila, Algeria)
- Dr. Amel Meliouh (U. of M'sila, Algeria)
- Dr. Belkacem Brahimi (U. M'sila, Algeria)
- Dr. Bilal Lounnas (U. M'sila, Algeria)
- Dr. Fares Mezrag (U. M'sila, Algeria)
- Dr. Farida Bouarab (U. Tizi Ouzou, Algeria)
- Dr. Fatima Boumahdi (U. Bida, Algeria)
- Dr. Fatma Salem (U. Tobruk, Libya)
- Dr. Fodil Belhadj (U. BBA, Algérie)
- Dr. Hassina Aliane (CERIST, Algeria)
- Dr. Kamel Amroun (U. Béjaia, Algeria)
- Dr. Koh Tieng Wei (U. Putra, Malaysia)
- Dr. Lyamine Guezouli (U. Batna 2, Algeria)
- Dr. Lyes Khelladi (CERIST, Algeria)
- Dr. Mahmoud Brahimi (U. M'sila, Algeria)
- Dr. Makhoulf Benazi (U. M'sila, Algeria)
- Dr. Meriem Laifa (U. BBA, Algérie)
- Dr. Meryem Regouid (U. Setif, Algeria)
- Dr. Mohamed Benouis (U. M'sila, Algeria)
- Dr. Mohamed Bounif (U. M'sila, Algeria)
- Dr. Mohamed Brahimi (The National Higher School of Artificial Intelligence (ENSIA))
- Dr. Mohamed Sahraoui (U. M'sila, Algeria)
- Dr. Mohammed Kamel Benkaddour (U. Ouargla, Algeria)
- Dr. Nacereddine Mouhoub (U. M'sila, Algeria)
- Dr. Nawal Sad Houari (U. Oran, Algeria)
- Dr. Nouredine Amraoui (U. M'sila, Algeria)
- Dr. Nouredine Chaib (U. Laghouat, Algeria)
- Dr. Ouarda Assas (U. Batna 2, Algeria)
- Dr. Rahima Bentercia (U. M'sila, Algeria)
- Dr. Said Kadri (U. of M'sila, Algeria)
- Dr. Tahar Mehenni (U. M'sila, Algeria)
- Dr. Abbas Cheddad (Belkinge Institute of Technology, Sweden)
- Dr. Abdelkader Laouid (U. El-Oued, Algeria)
- Dr. Mohamed Boubakir (U. Jijel, Algeria)
- Dr. Khedidja Derdour (U. M'sila, Algeria)

TABLE OF CONTENT		
Keynote Session		
Page No	Title	Authors
1	Metaverse and XR Immersive Technologies	Rachid GHERBI
2	Towards a More Powerful Semantics: Ontologies and Semantic Web for Industrial Transformation	Younes DJAGHLOUL
3	The Surge of Interest in M2M: Current State, Challenges and Perspectives	Azzedine BILAMI
Session 1: Innovations in Healthcare Systems		
Page No	Title	Authors
4	A new Authentication Protocol for RFID-based healthcare Applications	Anwar Nouredine Bahache, Chikouche Nouredine , Mohamed Bahache, Fares Mezrag
16	Using ontologies and web service composition to improve interoperability in the Algerian healthcare	KHELIL Hatem , BRAHIMI Mahmoud, ADDA Mo , MCHEICK Hamid
Session 2: Environmental Forecasting and Risk Modelling		
Page No	Title	Authors
22	Interval Markov chain modeling for weather forecasting	Salah Eddine SEMATI , Abdelkader GASMI
33	Fuzzy logic approach for forest fire risk estimation in Algeria	Harbouche Khadidja , Khentout Chabane , Haddad Dhai Eddine
Session 3: AI and Machine Learning Applications		
Page No	Title	Authors
41	Leveraging Transfer Learning and Data Augmentation for Enhanced Mental Health Systems in Chronic Disease Management.	Embarki Omar, Laifa Merieme , Mohdeb Djamila
49	Application of Deep Learning for the Automated Detection and Classification of Diabetic Retinopathy from High-Resolution Medical Images.	Ghizlane Khababa, Abdallah Khababa , Saad HAROUS , Mohamed Abdessamad KRACHE , Chouaib MAIZA
64	Simplifying the DBSCAN Algorithm with a Single-Parameter Approach.	Makhlouf Benazi
Session 4: Intelligent Systems and Automation		
Page No	Title	Authors
71	Modernizing Assessment: Overview of Computerized Adaptive Testing and AI.	Haddad Dhai Eddine, Khentout Chabane, Harbouche Khadidja
79	Deep Reinforcement Learning for Multi-intersection Traffic Signal Control: A New Cooperative Approach.	Tarek Amine Haddad
88	Empowering Customer Service with AI: A Multilingual Chatbot for Mobile Companies.	Tahar Mehenni

Session 5: Data Science, Computer Vision, and Image Processing		
Page No	Title	Authors
96	Pi-Based Scrambling Algorithm for Secure Image Watermarking.	Nesrine Benmeheni, Abdelhamid Benhocine, Nasser Eddine Mouhoub
104	Parallel Association Rules Mining Using GPUs and Reptile Search Algorithm.	Abderrahim Boukhalat, KamelEddine Heraguemi, Mohamed Benouis ,Brahim Bouderah and Samir Akhrout
120	A Comprehensive Evaluation of YOLOv9, YOLOv10, and YOLOv11 in Crowd Counting Applications.	Merwan Grabsi, Kameleddine Heraguemi
Session 6: Data Integration, Forecasting, and Recommender Systems		
Page No	Title	Authors
132	Real-Time Data Integration for Effective Congestion Forecasting.	Messaoud Bechere , Abdelbasset Barkat
141	A Concise Review of Online Benchmark Datasets for Recommender Systems.	Amna Meddeb, Lotfi Ben Romdhane
149	Trust in Normative Multi-agent Systems: Survey.	Guerfi Imane, Toufik Marir, Maroua Bouzid
Poster Session		
Page No	Title	Authors
157	Web Clicks Analysis by Episode Mining	Ouarem Oualid, Nouioua Farid, Mekhoukh Iamis , Hebboul Ahlem
158	Personality trait- driven ChatGPT adoption	Rima Benfredj , Farid Nouioua , Abderraouf Bouziane
159	A new heuristic for the flow-shop scheduling problem	Khaoula Djemla, Abdelhamid Benhocine, Nasser Eddine Mouhoub

KEYNOTE SESSION



Keynote 1

Metaverse and XR Immersive Technologies

By: **Pr. Rachid GHERBI**

Paris-Saclay University, France



Rachid Gherbi is a professor of computer science at the University of Paris-Saclay since 1988, specializing in the metaverse, virtual and augmented reality, and intelligent technologies. A graduate of USTHB, he earned his Master's and PhD in computer vision in 1992, followed by an HDR in 2001 on Human-Computer Interaction. He has led research groups at LIMSI CNRS and Genopole, supervised numerous Ph.D. and Master's students, and published over 100 international papers. Actively involved in European and national projects, he also contributes as an expert for MESRS in Algeria and in civil society. He currently heads the CU-PUIO at Paris-Saclay University and co-chairs the Science&Media club.

ABSTRACT

The Metaverse includes a range of immersive technologies like virtual reality and augmented reality. These tools are recognized for their ability to create engaging and immersive 3D experiences. They are being utilized in various industries, especially in healthcare.

In this presentation, I will provide an overview of the fundamental concepts behind these immersive technologies, the methods involved in their development and implementation, and their applications in different areas of healthcare and education.

Keynote 2

Towards a More Powerful Semantics: Ontologies and Semantic Web for Industrial Transformation

By: Dr. Younes DJAGHLOUL

ARHS Cube, Luxembourg



Younes Djaghoul received the Ph.D. degree in 2005. Following his doctoral studies, he worked as a Lecturer at the University of M'sila for one year. In 2009, he joined the research centers Henri Tudor and LIST in Luxembourg as a Research and Development Engineer. During his time there (2009–2014), he focused on semantic web technologies, contributing to projects on recommender systems, adaptive content delivery, e-learning platforms, and e-government solutions.

In 2016, Dr. Djaghoul transitioned to the industry, joining ARHS Cube, a renowned software company. As a Senior IT Consultant and Software Engineer Expert, he delivered complex, data-intensive projects. His expertise included utilizing graph databases, ontology-based solutions, and advanced AI technologies, as well as employing modern software engineering practices.

ABSTRACT

The use of ontologies has become increasingly beneficial in industrial contexts, transcending the boundaries of purely academic research. However, transitioning the outcomes of research in this field to practical applications remains a significant challenge. Key questions arise in this context: Does employing more complex modeling languages effectively address the challenges of knowledge representation? Will the community adopt and sustain the use of a given ontology? Who ensures its evolution post-deployment? How can various versions be evaluated to select the most appropriate one? Furthermore, even the most mature ontologies often face obstacles when directly applied in real-world systems.

This study leverages our experience over the past decades across various projects to examine the practical challenges of developing ontology-based applications. Specifically, Through real-world examples and case studies—including both successes and less successful attempts—we analyze critical issues such as scalability, inconsistent modeling practices, integration difficulties, and the adoption and maintenance of ontologies in operational environments. By addressing these challenges, the study aims to provide actionable insights into bridging the gap between theoretical research and the deployment of reliable, functional ontology-based systems.

Keynote 3

The Surge of Interest in M2M: Current State, Challenges and Perspectives

By: Pr. Azzedine BILAMI
Batna2 University, Algeria



Azeddine Bilami received the Ph.D. degree in 2005. He is currently serving as a Professor (full position) at the department of computer science, and the Head of LaSTIC laboratory, at University of Batna2. Prof. A. Bilami has more than fifty publications in international journals and conferences including IJCA (Actapress), IJSNet (INDERSCIENCE), IEEE Communications Letters, Springer Verlag, IGI Global, and Elsevier publications. He also served as a member of editor boards, a member of steering committees and a member of technical committees in many international conferences and journals (COMNET, COMCOM (Elsevier), HINDAWI journals, KSII Transactions on Internet and Information Systems, IAJIT, Journal: Journal of King Saud University, NGNS..). His current research interests are mainly focused on Network Security, Routing Protocols, Mobility, QoE and QoS in wireless and mobile networks, WSNs and the IoT(Internet of Things).

ABSTRACT

Machine-to-Machine (M2M) technology enables devices to communicate and exchange data autonomously, serving as a keystone of the Internet of Things (IoT). Since its origins in the 1990s with early telemetry applications, M2M has grown significantly in recent years, powered by the explosion of connected devices and the advent of technologies like 5G.

This presentation will trace the key indicators in M2M's evolution and explore its current applications across critical sectors such as industry, healthcare, transportation, and energy. It will also address challenges such as energy efficiency, scalability, and data security, with a particular focus on gainful blockchain technologies to ensure the integrity and traceability of communications. Special attention will be given to lightweight protocols designed for the constrained resources of M2M devices.

The talk will also highlight emerging technologies such as 5G networks, edge computing, and artificial intelligence, examining their role in optimizing interactions between autonomous machines.

Finally, this presentation will outline the future prospects for M2M communications as they continue to evolve, driven by emerging trends and innovations.

SESSION 1

Innovations in Healthcare Systems



A new Authentication Protocol for RFID-based healthcare Application

Anwar Nouredine Bahache¹, Chikouche Nouredine², Mohamed Bahache², Fares Mezrag²

¹ University of M'sila, Computer science department, M'sila, Algeria

anwar.bahache@gmail.com 1

² LIAM Laboratory, University of M'sila, Computer science department, M'sila, Algeria

nouredine.chikouche@univ-msila.dz 2

mohamed.bahache@univ-msila.dz 3

fares.mezrag@univ-msila.dz 4

Abstract

With the rise of Internet and Communication Technologies (ICT), RFID (Radio Frequency Identification) technology has become indispensable in healthcare for tracking and managing tagged medical devices. Widely deployed across numerous healthcare assets, RFID tags face critical security and privacy challenges, as adversaries can eavesdrop, alter, or intercept transmitted data, compromising the confidentiality of healthcare personnel and patients. Although various lightweight RFID authentication schemes have been proposed to address these risks, many remain vulnerable to attacks, including replay, impersonation, and de-synchronization. To overcome these security limitations, we propose a resilient and efficient RFID authentication scheme tailored for IoT-enabled healthcare applications. Leveraging Elliptic Curve Cryptography (ECC) to ensure high security with minimal computational load, our scheme demonstrates strong resistance to all assessed attack vectors. Rigorous security and performance evaluations confirm that our protocol achieves an optimal balance of security and efficiency, positioning it as a practical and secure solution for real-time healthcare environments.

Keywords: RFID systems, authentication protocols, healthcare applications, Elliptic curve cryptography, security

1. Introduction

With rapid advancements in ICT and automated medication systems, RFID (Radio Frequency Identification) and WBAN (Wireless Body Area Network) technologies are increasingly adopted in healthcare to enhance patient safety [1]. As a key technology in pervasive computing, RFID enables unique, simultaneous identification of multiple objects over a channel [2]. RFID has numerous applications, including automatic payment, access control, toll collection, personnel identification, e-healthcare, and supply chain management [3, 4].

In healthcare, RFID offers benefits like theft prevention, reduced human labor, improved productivity, and cost efficiency [5, 6]. Emerging smart healthcare systems leverage RFID for monitoring, mobility, and remote access to patient information via cloud servers [7]. However, patient misidentification remains a concern; RFID mitigates such risks by supporting asset and patient tracking, boosting safety, and optimizing operational efficiency [8]. Despite these benefits, security, privacy, and human safety concerns remain barriers to adoption [9].

RFID systems comprise three main components: tags, readers, and backend servers. Tags are embedded microchips, categorized as passive, active, or semi-active [10]. They operate across low (LF), high (HF), and ultra-high frequencies (UHF) with varying ranges. Readers communicate with tags to gather data, which is stored on backend servers, often replaced in healthcare

by cloud servers for cost efficiency and scalability [11].

In healthcare, secure channels are used between readers and servers, while insecure wireless channels connect tags (e.g., patients) to readers (e.g., medical staff) [12]. These vulnerabilities necessitate a robust RFID authentication scheme to safeguard patient data and medical records. Key objectives of this scheme include:

- Ensuring mutual authentication between tag, reader, and server.
- Meeting security requirements for RFID systems.
- Providing resistance against known attacks.
- Minimizing computational and storage costs.

The remainder of our paper is organized as follows: Section II presents existing related works in literature. In Section III we present the proposed protocol with its preliminaries. In Section IV a security analysis is presented followed by a performance analysis in Section V, with a discussion. Finally, our paper is concluded in Section VI.

2. Related work

In this section, some related works in the field are introduced. In recent years, various RFID authentication schemes have been introduced to secure RFID systems against diverse security threats [13]. Low-cost RFID systems face challenges in ensuring complete security and privacy due to insecure communication between tags and readers. To address these issues, we review previous schemes along with their cryptographic methods, strengths, and weaknesses.

Wu et al. [12] introduced a lightweight RFID scheme for e-healthcare that uses only hash functions to meet core security needs (mutual authentication, integrity, anonymity) while maintaining low-cost operation suited for low-spec devices. Baashirah et al. [14] addressed sensitive data security in healthcare RFID systems by refining Zhang and Qi's [15] protocol. They used ECC for enhanced performance and added hash functions to protect tag identifiers, achieving mutual authentication with increased security but additional computation.

Zhou et al. [16] developed a privacy-preserving RFID protocol for healthcare, leveraging hash and quadratic residues to enhance security while minimizing device computation. This approach addresses replay attacks, integrity, and untraceability in response to flaws identified in Li et al. [17] work. Lv et al. [18] introduced a fully anonymous ECC-based authentication scheme in 2020, emphasizing security against both outsider and insider attacks. Using bilinear pairing and a certification authority, they optimized device efficiency and demonstrated the protocol's scalability in healthcare applications, though the approach increased computational costs. Noori et al. [19] presented a scalable, efficient ECC and hash-based protocol for healthcare, allowing low-cost addition or revocation of devices and focusing on secure, scalable RFID communication.

Zhu Feng [20] critiqued Safkhani and Vasilakos's protocol [21] and proposed a secure RFID protocol based on hash and square root operations. Although effective in privacy, its high resource demand highlights a trade-off between security and performance. Xie et al. [22] addressed back-end server vulnerabilities in RFID by incorporating an indistinguishability obfuscation technique. Extending to cloud storage, they reduced on-device costs and mitigated data leakage risks associated with traditional servers. Salem and Amin [23] designed a privacy-preserving protocol for Telecare Medicine Information Systems (TMIS) using El-Gamal cryptography to safeguard patient safety.

Lately, Agrahari and Varma [24] applied ECC-based Qu-Vanstone certificates for mobile, secure, and scalable healthcare authentication with minimal computation and key size requirements.

Izza et al. [25] proposed an ECC and ECDSMR-based RFID protocol to improve Naeem et al. [26] scheme for wearable healthcare networks (WBANs), focusing on strengthened security over the internet. Other noteworthy protocols include Xie et al. [27] used a VPN to create secure backend channels for a cloud-based RFID authentication protocol, organizing data in an encrypted hash table and using operators like XOR, hash functions, and symmetric encryption. Their protocol achieved mutual authentication and tag privacy but was later shown to have vulnerabilities, including susceptibility to location tracking [28]. Xiao et al. [29] developed cloud-RAPIC, which facilitates secure RFID authentication without third-party intermediaries. It safeguards mutual authentication, forward security, data integrity, and anonymity. However, it too was found vulnerable to certain attacks, including reader impersonation and tag privacy breaches [28]. Zhenguo Zhao [30] proposed an ECC-based secure RFID scheme for healthcare systems that protects against forward untraceability. Yet, Farash et al. [31] identified weaknesses in forward security.

Rahman et al. [13] introduced a privacy-preserving RFID framework for healthcare environments, emphasizing privacy in tracking and access control for healthcare data. The framework also addresses privacy-scalability trade-offs. Fan et al. [32] proposed a lightweight RFID scheme for IoT healthcare, enhancing privacy and offering protection against common attacks like DoS and replay. However, subsequent research [33] highlighted issues with tag anonymity and traceability, prompting improvements in Aghili et al.'s bSecLAP scheme. Additional RFID schemes focus on cloud-based healthcare systems (Fan et al. [34]) and database protection from sensitive information leakage, employing a range of cryptographic methods. Song et al. [35] introduced ZKAP, a zero-knowledge RFID authentication protocol, offering strong privacy features but lacking formal security verification.

3. Proposed protocol

In this section, we present our proposed protocol. In addition to a basic system model for RFID-based healthcare systems and a security model that describes the adversary's capabilities.

3.1. System model

The proposed RFID-based healthcare system architecture consists of several key components: RFID tags, an RFID reader, and a Trusted Authority (TA) responsible for registrations and management of the system. Figure 1 represents our basic RFID architecture model.

3.1.1. System architecture:

- *RFID Tags*: Each patient or healthcare entity is assigned an RFID tag that contains unique identification information. These tags can store various types of data, including patient medical history, allergies, medications, and other relevant health information. The tags are passive or active, depending on their design, and are used for quick identification and data retrieval.
- *RFID Reader*: The RFID reader is a device that emits RF signals to communicate with RFID tags. It can read and write data to the tags within its range. In the healthcare system, the reader is typically placed at strategic locations, such as at the entrance of healthcare facilities or at nursing stations, to facilitate the quick retrieval of patient information.
- *Trusted Authority (TA)*: The Trusted authority is responsible for managing the registration and authentication processes within the RFID-based healthcare system. It ensures that only authorized personnel can access sensitive patient information. The TA maintains a secure database of registered users, including healthcare professionals and patients, and manages

their respective permissions and roles. It also connects directly to the reader to retrieve the tags' information.

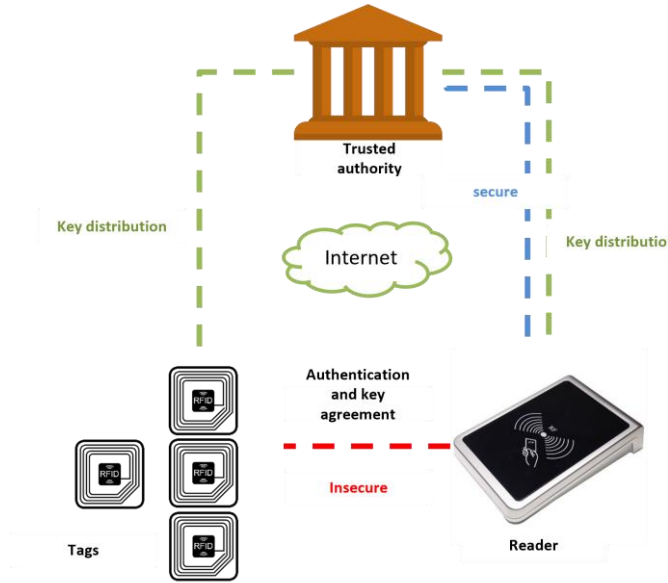


Figure 1. Our basic model for RFID systems.

3.1.2. System Operations:

The operation of our RFID-based healthcare system/architecture can be described as follows:

- *Registration:* When a new patient is admitted, the TA registers the patient in the system. This process involves issuing a unique RFID tag to the patient and storing their relevant information in the secure database.
- *Data Retrieval:* When a healthcare professional needs to access a patient's information, they use the RFID reader to scan the patient's RFID tag which represents a direct communication to the T A. The reader retrieves the associated data from the T A's database, allowing for quick and efficient access to the patient's medical history.
- *Access Control:* The TA enforces strict access control policies to ensure that only authorized personnel can access sensitive data. This includes authentication mechanisms to verify the identity of healthcare professionals before granting access to patient information.
- *Data Security:* To protect patient information from unauthorized access and ensure data integrity, the system employs various security measures, including ECC, hash, secure communication protocols, and regular security audits.

3.2. Security model

It is essential to model intruder behavior to counter network attacks to predict and prevent potential threats effectively. Attackers may act passively, simply eavesdropping on transmissions between legitimate entities, or actively by intercepting and corrupting messages. This work employs the robust Dolev–Yao (DY) threat model [36], which assumes the intruder has maximal capabilities, challenging the protocol's resistance [37]. Despite appearing unrealistic, the DY model enhances IoT scheme resilience and defense against various attacks [38].

3.3. Protocol description

Our proposed protocol includes two primary phases: initialization and registration, which may be further split, and authentication.

3.3.1. Initialization and Registration Phase

In IoT-based healthcare systems, a Trusted Authority (TA) generates system parameters and ensures secure registration for all participating entities.

- Setup by TA:
 - Elliptic curve selection: the TA selects an elliptic curve E_q over a finite field F_q , where q is a prime number.
 - Base point: the base point P of order n is chosen on E_q , and this is shared across the network.
 - TA key generation:
 - TA generates a private key α , a random integer in $[1, n - 1]$.
 - TA computes its public key as:

$$TA_{pk} = \alpha \cdot P \quad (1)$$

- Public and private key generation for each entity U_v (e.g., reader U_R and tag U_S):
 - Each entity U_v generates a random integer $c_v \in [1, n - 1]$ as its temporary private value.
 - The entity computes:

$$d_v = c_v \cdot P \quad (2)$$

- U_v sends (d_v, ID_v) to the TA, where ID_v is the unique identifier of U_v .
- Key assignment by TA for each U_v : upon receiving d_v and ID_v from U_v , the TA:
 - Selects a random integer $w_v \in [1, n - 1]$.
 - Computes:

$$y_v = w_v \cdot P + d_v \quad (3)$$

$$z_v = w_v + ((y_v)x + ID_v) \cdot \alpha \mod n \quad (4)$$

- The TA returns (y_v, z_v) to U_v .
- Computation and verification of the secret key by U_v : upon receiving (y_v, z_v) , U_v calculates its private key as:

$$x_v = z_v + c_v \mod n \quad (5)$$

Verification: U_v verifies its private key by checking:

$$x_v \cdot P = y_v + ((y_v)x + ID_v) \cdot TA_{pk} \quad (6)$$

If the verification holds, U_v confirms the validity of x_v .

- Distribution of keys: each entity now holds its private key x_v and public key y_v . Specifically, U_R (the reader) has x_R and y_R , and U_S (the server or tag) has x_S and y_S .

3.3.2. Authentication Phase

- Preliminaries each entity (reader U_R and tag U_S) has a private key (x_R for reader, x_S for tag) and corresponding public key ($y_R = x_R \cdot P$ for reader, $y_S = x_S \cdot P$ for tag). Both entities share a secure hash function $H(\cdot)$ and a synchronized lock with tolerance Δt .
- Authentication steps with concealed ID_s using scalar multiplication and timestamps

- Step 1: reader initiates authentication the reader U_R generates a random nonce N_R and a timestamp T_R . The reader computes a temporary identifier ID'_R using its private key x_R , the tag's public key y_S , and timestamp T_R :

$$ID'_R = H(x_R \cdot y_S \parallel T_R) \quad (7)$$

The reader sends (ID'_R, N_R, T_R) to the tag U_S :

$$U_R \rightarrow U_S : ID'_R, N_R, T_R$$

- Step 2: tag verifies timestamp and responds with verification message. Upon receiving ID'_R , N_R , and T_R , the tag U_S checks T_R to ensure it is within an acceptable range (i.e., $|T_R - T_{current}| \leq \Delta t$). If T_R is valid, the tag generates its own nonce N_S and timestamp T_S . The tag computes a shared value K_{RS} as:

$$K_{RS} = H(x_S \cdot y_R + N_R + N_S) \quad (8)$$

The tag computes a temporary identifier for itself, ID'_S , using its private key x_S , the reader's public key y_R , and timestamp T_S :

$$ID'_S = H(x_S \cdot y_R \parallel T_S) \quad (9)$$

The tag then computes an authentication message $Auth_{Tag}$:

$$Auth_{Tag} = H(K_{RS} \parallel N_R \parallel ID'_S) \quad (10)$$

The tag sends $(Auth_{Tag}, N_S, T_S, ID'_S)$ to the Reader:

$$U_S \rightarrow U_R : Auth_{Tag}, N_S, T_S, ID'_S$$

- Step 3: reader verifies tag's authenticity and timestamp upon receiving $Auth_{Tag}$, N_S , T_S , and ID'_S , the reader checks the timestamp T_S to ensure it is within the acceptable range. If T_S is valid, the Reader calculates K_{RS} as:

$$K_{RS} = H(x_R \cdot y_S + N_R + N_S) \quad (11)$$

The reader computes $Auth'_{Tag} = H(K_{RS} \parallel N_R \parallel ID'_S)$ and verifies that $Auth'_{Tag} = Auth_{Tag}$. If they match, the tag is authenticated.

- Step 4: reader sends authentication response to confirm its identity to the tag, the reader computes an authentication response $Auth_{Reader}$:

$$Auth_{Reader} = H(K_{RS} \parallel N_S \parallel ID'_R) \quad (12)$$

The reader sends $(Auth_{Reader}, T_R)$ to the Tag:

$$U_R \rightarrow U_S : Auth_{Reader}, T_R$$

- Step 5: tag Verifies reader's authenticity and timestamp. Upon receiving $Auth_{Reader}$ and T_R , the tag verifies T_R to ensure it is recent. If the timestamp is valid, the tag computes $Auth'_{Reader} = H(K_{RS} \parallel N_S \parallel ID'_R)$ and checks if $Auth'_{Reader} = Auth_{Reader}$. If they match, the reader is authenticated.

- Session key establishment after successful mutual authentication, both reader and tag can establish a session key SK for secure communication based on the shared secret K_{RS} :

$$SK = H(K_{RS} \parallel N_R \parallel N_S \parallel T_R \parallel T_S) \quad (13)$$

The summary of the mutual authentication protocol is as follow:

- Message 1: $U_R \rightarrow U_S : ID'_R, N_R, T_R$
- Message 2: $U_S \rightarrow U_R : Auth_{Tag}, N_S, T_S, ID'_S$
- Verification by reader: reader checks $Auth'_{Tag} = Auth_{Tag}$ and verifies T_S
- Message 3: $U_R \rightarrow U_S : Auth_{Reader}, T_R$
- Verification by tag: tag checks $Auth'_{Reader} = Auth_{Reader}$ and verifies T_R

4. Security analysis

This section presents an informal security analysis for the proposed protocol. In addition, to a detailed security comparison with different protocols.

4.1. Informal analysis

- 4.1.1. *Confidentiality of Identifiers*: By utilizing hash functions and scalar multiplication with private and public keys, the protocol generates temporary identifiers (ID'_R and ID'_S), effectively concealing the actual identifiers of the Reader and Tag from potential adversaries.
- 4.1.2. *Integrity of Messages*: Authentication messages ($Auth_{Tag}$ and $Auth_{Reader}$), based on a shared secret K_{RS} along with nonces and temporary identifiers, are included in each communication. This ensures message integrity, as any unauthorized alteration can be detected.
- 4.1.3. *Mutual Authentication*: The protocol enforces mutual authentication by requiring both the reader and tag to verify each other through matching authentication messages, thus preventing impersonation attacks.
- 4.1.4. *Replay Attack Resistance*: Nonces and timestamps prevent replay attacks by ensuring message freshness. Any replayed message would fail timestamp validation, blocking unauthorized attempts.
- 4.1.5. *Forward Secrecy*: Temporary identifiers and session keys, generated from nonces and timestamps, provide forward secrecy. Compromise of a private key does not affect past session security due to context-specific key derivation.
- 4.1.6. *Resistance to Man-in-the-Middle (MitM) Attacks*: The protocol resists MitM attacks by authenticating exchanged messages, making it impossible for an attacker to forge valid messages without access to the shared secret or current nonces.
- 4.1.7. *Key Management*: Elliptic curve cryptography (ECC) strengthens key management by delivering high security with small key sizes, enhancing efficiency and lowering key compromise risk.

4.1.8. *Time Synchronization Issues*: Timestamps improve security but depend on synchronized clocks between the reader and tag. Periodic synchronization can mitigate potential failures due to desynchronized clocks.

4.1.9. *Hash Function Security*: The protocol's security relies on a strong hash function $H(\cdot)$ (e.g., *SHA-256*) to prevent vulnerabilities, such as collision attacks, which would undermine the protocol's integrity.

4.2. Security comparison

Security is paramount in RFID authentication. Table I summarizes the possible attacks and security requirements addressed or unmet by each studied protocol. These attacks, as shown in Table I, were identified by various authors who have since proposed enhancements or alternative schemes.

TABLE I
DIFFERENT SECURITY REQUIREMENTS AND ATTACKS IN THE STUDIED PROTOCOLS

Protocol	A1	A2	A3	A4	A5	A6	A7
[14]	✓	*	✓	✓	✓	✓	*
[16]	✓	✓	✓	✓	*	×	✓
[18]	✓	✓	✓	✓	*	*	*
[19]	✓	✓	*	✓	*	*	✓
[23]	✓	✓	✓	✓	✓	✓	✓
[24]	✓	*	✓	✓	*	*	✓
[25]	×	×	×	✓	*	×	✓
Our	✓	✓	✓	✓	✓	✓	✓

✓: Ensure/Resist ×: Fails to ensure/resist *: not discussed

A1: Anonymity A2: Forward/Backward secrecy A3: Untraceability

A4: Replay attack A5: DoS attack A6: Desynchronization attack

A7: Impersonation attack

The comparison reveals that most protocols do not fully satisfy all essential security and privacy requirements. However, our proposed protocol not only meets all these requirements but also resists every discussed attack, ensuring robust protection across all fronts. This makes our protocol a superior choice for secure, resilient RFID authentication in vulnerable environments.

5. Performance analysis

Table II presents a performance comparison of the studied protocols, where computational and storage costs are evaluated for the tags. Given the limited computational power and memory capacity of tags, practical applications must prioritize minimizing computation, communication, and storage requirements.

TABLE II
EVALUATION OF THE STUDIED PROTOCOLS

Protocol	Computational cost	Communication cost	Storage cost
[14]	$4T_H + 2T_{ECM} + T_{ECA}$	$3L + 2L_{ECM}$	$3L_{ECM}$
[16]	$4T_H + 3T_{MODS}$	$9L_B + 7L + 2L_{TS}$	$L + L_B + L_{ID}$
[18]	$6T_{ECM} + 2T_{MULT} + 2T_{MOD} + T_H$	$2L + 5L_{ECM} + L_B$	$4L_{ECM} + 4L$
[19]	$T_{ECM} + 2T_H + 2T_S$	$4L_B$	$L_{ECM} + 4L$
[23]	$3T_H + 2T_{MODS} + T_{MULT}$	$3L + 2L_B$	$3L + L_B$
[24]	$2T_{ECM} + T_{ECA} + 2T_H + 3T_{MULT}$	$2L_{ECM} + L_{ID} + 2L$	$L_{ECM} + 6L + L_{ID}$
[25]	$3T_{ECM} + 6T_H + T_S$	$6L + 3L_{ECM} + 5L_{TS}$	$L_{ID} + 2L + 3L_{ECM} + L_B$
Our	$2T_{ECM} + 3T_H$	$5L + 2L_{TS}$	$2L_{ECM} + L_{ID}$

The experiments were run 50 times on 1024-bit messages in an Ubuntu terminal on a Windows 10 machine, using a Ryzen 5 3600x 3.8GHz CPU and 16 GB 3200 MHz RAM. For symmetric and asymmetric cryptographic operations, we utilized the OpenSSL library.

Table III lists the cryptographic primitives employed in each protocol, along with their computational times. Note that generating a 256-bit random number (RNG) takes approximately the same time as computing a 256-bit hash function (SHA-256). Also, we define assumptions to calculate the communication and storage cost as shown in table IV.

TABLE III
IMPLEMENTATION RESULTS

Operation	Notation	Timing (in μs)
Hash function/RNG (SHA-256)	T_H	184.56
Symmetric Enc/Decryption (AES-256)	T_S	37
Scalar point multiplication (SECG-P256/ 256-bit prime field)	T_{ECM}	1177
ECC Addition (SECG-P256/ 256-bit prime field)	T_{ECA}	553.68
Modular operation	T_{MOD}	1752
Modular square/exponentiation operation	T_{MODS}	1896.4
Modular multiplication multiplications	T_{MULT}	144.4

Figure 2 presents the computational, communication, and storage costs for the studied protocols on the tag side. Below, we analyze the performance based on these metrics.

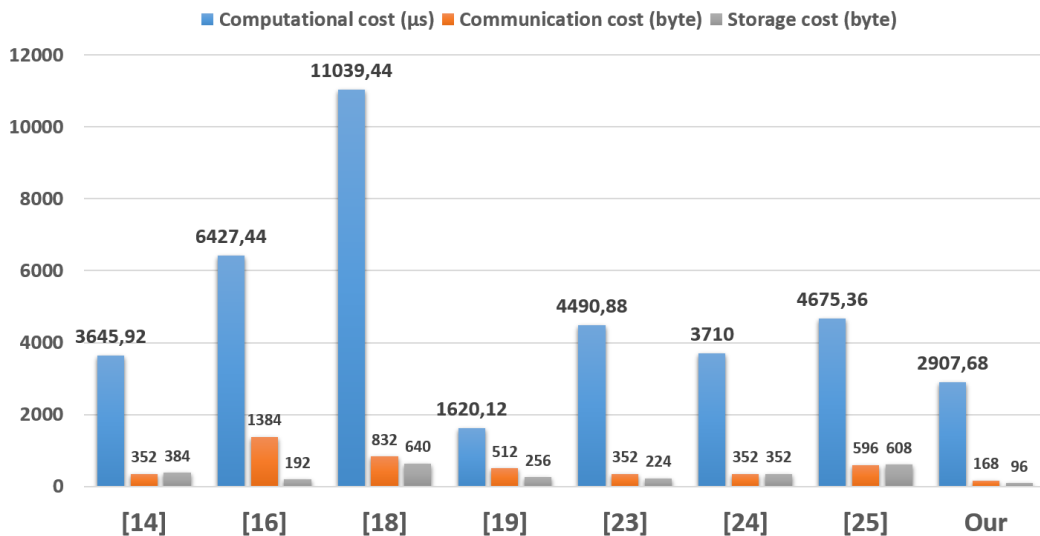


Figure 2. Costs Comparison

5.1. Computational Cost:

Noori et al. [19] and Salem et al. [23] protocols have the lowest computational costs among existing schemes, with 1620.12 μ s and 2450.08 μ s, respectively. However, protocols such as Zhou et al. [16] and Lv et al. [18] impose significantly higher computational demands (6427.44 μ s and 11039.36 μ s, respectively) due to the use of ECC and quadratic residue operations. In contrast, our proposed protocol achieves a more balanced approach, taking only 2907.68 μ s, maintaining strong security without the high computational burden associated with extensive ECC operations.

5.2. Communication Cost:

The communication costs of Baashirah et al. [14], Salem et al. [23], and Agrahari et al. [24] protocols are among the lowest, at 352 bytes, while protocols like Zhou et al. [16] and Lv et al. [18] incur much higher costs, up to 1384 bytes. Our protocol keeps communication overhead minimal, with a cost of 168 bytes, showcasing efficient data exchange without compromising on security.

5.3. Storage Cost:

In terms of storage, our protocol matches its communication and computation efficiency, requiring only 96 bytes of storage. By minimizing storage requirements, our approach supports resource-constrained RFID systems more effectively. In summary, our protocol offers an optimal balance, achieving enhanced security and performance with reduced computational, communication, and storage demands, positioning it as a highly practical solution for RFID applications in constrained environments.

6. Conclusion

In conclusion, this research introduced a lightweight RFID authentication protocol leveraging ECC, tailored for the unique security and efficiency demands of IoT-based healthcare systems. By effectively mitigating risks like replay, impersonation, and desynchronization attacks, our protocol demonstrates superior resilience over existing schemes. The performance analysis shows that it maintains low computational and storage costs, ensuring practicality for resource-limited RFID tags in real-time healthcare applications. Looking ahead, this protocol lays the groundwork for scalable, secure RFID solutions that can adapt to evolving IoT standards and expanding healthcare networks. Future research could focus on optimizing the scheme for broader, large-scale healthcare deployments, further strengthening privacy and reliability in an increasingly interconnected medical landscape.

References

- [1] A. N. Bahache, N. Chikouche, and F. Mezrag, "Authentication schemes for healthcare applications using wireless medical sensor networks: A survey," *SN Computer Science*, vol. 3, no. 5, p. 382, 2022.
- [2] R. Want, "An introduction to RFID technology," *IEEE pervasive computing*, vol. 5, no. 1, pp. 25–33, 2006.
- [3] B. Nath, F. Reynolds, and R. Want, "RFID technology and applications," *IEEE Pervasive computing*, vol. 5, no. 1, pp. 22–24, 2006.
- [4] N. Dinarvand and H. Barati, "An efficient and secure RFID authentication protocol using elliptic curve cryptography," *Wireless Networks*, vol. 25, no. 1, pp. 415–428, 2019.
- [5] A. N. Bahache and N. Chikouche, "A comparative analysis of RFID authentication protocols for healthcare applications," in *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP)*. IEEE, 2021, pp. 1–6.
- [6] W. Yao, C.-H. Chu, and Z. Li, "The adoption and implementation of RFID technologies in healthcare: a literature review," *Journal of medical systems*, vol. 36, pp. 3507–3525, 2012.
- [7] H.-Y. Chien, C.-C. Yang, T.-C. Wu, and C.-F. Lee, "Two RFID-based solutions to enhance inpatient medication safety," *Journal of Medical Systems*, vol. 35, pp. 369–375, 2011.

- [8] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.
- [9] W. Yao, C.-H. Chu, and Z. Li, "The use of RFID in healthcare: Benefits and barriers," in 2010 IEEE International Conference on RFID-Technology and Applications. IEEE, 2010, pp. 128–134.
- [10] M. Shariq, K. Singh, P. K. Maurya, A. Ahmadian, and M. R. K. Ariffin, "Urasp: An ultralightweight RFID authentication scheme using permutation operation," *Peer-to-Peer Networking and Applications*, vol. 14, pp. 3737–3757, 2021.
- [11] M.-G. Avram, "Advantages and challenges of adopting cloud computing from an enterprise perspective," *Procedia Technology*, vol. 12, pp. 529–534, 2014.
- [12] F. Wu, L. Xu, S. Kumari, X. Li, A. K. Das, and J. Shen, "A lightweight and anonymous RFID tag authentication protocol with cloud assistance for e-healthcare applications," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, pp. 919–930, 2018.
- [13] F. Rahman, M. Z. A. Bhuiyan, and S. I. Ahamed, "A privacy preserving framework for RFID based healthcare systems," *Future generation computer systems*, vol. 72, pp. 339–352, 2017.
- [14] R. Baashirah, A. Abuzneid, A. Tammineedi, M. Mathew, N. Bandaru, and P. Delair, "Improve healthcare safety using hash-based authentication protocol for RFID system," in 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, 2018, pp. 800–805.
- [15] Z. Zhang and Q. Qi, "An efficient RFID authentication protocol to enhance patient medication safety using elliptic curve cryptography," *Journal of medical systems*, vol. 38, no. 5, pp. 1–7, 2014.
- [16] Z. Zhou, P. Wang, and Z. Li, "A quadratic residue-based RFID authentication protocol with enhanced security for tmis," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 9, pp. 3603–3615, 2019.
- [17] C.-T. Li, C.-Y. Weng, and C.-C. Lee, "A secure RFID tag authentication protocol with privacy preserving in telecare medicine information system," *Journal of medical systems*, vol. 39, no. 8, pp. 1–8, 2015.
- [18] J. Lv, N. Xi, and X. Rao, "A fully anonymous authentication scheme based on medical environment," in *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 2019, pp. 484–498.
- [19] D. Noori, H. Shakeri, and M. N. Torshiz, "Scalable, efficient, and secure RFID with elliptic curve cryptosystem for internet of things in healthcare environment," *EURASIP Journal on Information Security*, vol. 2020, no. 1, pp. 1–11, 2020.
- [20] F. Zhu, "Secmap: a secure RFID mutual authentication protocol for healthcare systems," *IEEE Access*, vol. 8, pp. 192 192–192 205, 2020.
- [21] M. Safkhani and A. Vasilakos, "A new secure authentication protocol for telecare medicine information system and smart campus," *IEEE Access*, vol. 7, pp. 23 514–23 526, 2019.
- [22] S. Xie, F. Zhang, and R. Cheng, "Security enhanced RFID authentication protocols for healthcare environment," *Wireless Personal Communications*, vol. 117, no. 1, pp. 71–86, 2021.
- [23] F. M. Salem and R. Amin, "A privacy-preserving RFID authentication protocol based on el-gamal cryptosystem for secure tmis," *Information Sciences*, vol. 527, pp. 382–393, 2020.
- [24] A. K. Agrahari and S. Varma, "A provably secure RFID authentication protocol based on ECQV for the medical internet of things," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1277–1289, 2021.
- [25] S. Izza, M. Benssalah, and K. Drouiche, "An enhanced scalable and secure RFID authentication protocol for wban within an iot environment," *Journal of Information Security and Applications*, vol. 58, p. 102705, 2021.
- [26] M. Naeem, S. A. Chaudhry, K. Mahmood, M. Karuppiah, and S. Kumari, "A scalable and secure RFID mutual authentication protocol using ecc for internet of things," *International Journal of Communication Systems*, vol. 33, no. 13, p. e3906, 2020.
- [27] W. Xie, L. Xie, C. Zhang, Q. Zhang, and C. Tang, "Cloud-based RFID authentication," in 2013 IEEE international conference on RFID (RFID). IEEE, 2013, pp. 168–175.
- [28] S. Abughazalah, K. Markantonakis, and K. Mayes, "Secure improved cloud-based RFID authentication protocol," in *International Workshop on Data Privacy Management*. Springer, 2014, pp. 147–164.

- [29] H. Xiao, A. A. Alshehri, and B. Christianson, "A cloud-based RFID authentication protocol with insecure communication channels," in 2016 IEEE Trustcom/BigDataSE/ISPA. IEEE, 2016, pp. 332–339.
- [30] Z. Zhao, "A secure RFID authentication protocol for healthcare environments using elliptic curve cryptosystem," *Journal of medical systems*, vol. 38, pp. 1–7, 2014.
- [31] M. S. Farash, O. Nawaz, K. Mahmood, S. A. Chaudhry, and M. K. Khan, "A provably secure RFID authentication protocol based on elliptic curve for healthcare environments," *Journal of medical systems*, vol. 40, pp. 1–7, 2016.
- [32] K. Fan, W. Jiang, H. Li, and Y. Yang, "Lightweight RFID protocol for medical privacy protection in iot," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1656–1665, 2018.
- [33] S. F. Aghili, H. Mala, P. Kaliyar, and M. Conti, "Seclap: Secure and lightweight RFID authentication protocol for medical iot," *Future Generation Computer Systems*, vol. 101, pp. 621–634, 2019.
- [34] K. Fan, S. Zhu, K. Zhang, H. Li, and Y. Yang, "A lightweight authentication scheme for cloud-based RFID healthcare systems," *IEEE Network*, vol. 33, no. 2, pp. 44–49, 2019.
- [35] J. Song, P.-W. Harn, K. Sakai, M.-T. Sun, and W.-S. Ku, "An RFID zero-knowledge authentication protocol based on quadratic residues," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 12 813–12 824, 2021.
- [36] D. Dolev and A. Yao, "On the security of public key protocols," *IEEE Transactions on information theory*, vol. 29, no. 2, pp. 198–208, 1983.
- [37] A. N. Bahache, N. Chikouche, and S. Akleylek, "Securing cloud-based healthcare applications with a quantum-resistant authentication and key agreement framework," *Internet of Things*, vol. 26, p. 101200, 2024.
- [38] M. Abomhara and G. M. Køien, "Security and privacy in the internet of things: Current status and open issues," in 2014 international conference on privacy and security in mobile systems (PRISMS). IEEE, 2014, pp. 1–8.

Using ontologies and web service composition to improve interoperability in the Algerian healthcare

KHELIL Hatem¹, BRAHIMI Mahmoud², ADDA Mo³, MCHEICK Hamid⁴

¹ Mohamed Boudiaf University, Computer Science Department(LLAM), M'sila, Algeria, hathem.khelil@univ-msila.dz.

² Mohamed Boudiaf University, Computer Science Department(LLAM), M'sila, Algeria, mahmoud.brahimi@univ-msila.dz.

³ The University of Portsmouth and Cambridge Education Group, School of Computing, Portsmouth, UK, mo.adda@port.ac.uk.

⁴ University of Quebec at Chicoutimi (UQAC), Department of Computer Science and Mathematics, Chicoutimi, Canada, hamid_mcheick@uqac.ca.

Abstract

Interoperability in healthcare presents a key concept that allows different entities in the healthcare sector to exchange patient medical records in order to coordinate patient care and monitoring. The Algerian healthcare sector offers an opportunity to adopt new IT systems that facilitate collaboration, information exchange and communication between systems in order to optimize development efforts, reduce costs and eliminate errors. In this paper, we explore the use of ontologies and web service composition as strategic approaches to overcome interoperability challenges, enhance data sharing, and promote collaborative healthcare practices.

Keywords: *Algeria healthcare, Composition, Interoperability, Ontology, Semantic Web services.*

1. Introduction

After independence in 1962, the Algerian health system witnessed a transformation from an insufficient health system, suffering from disparities in the quality of services and concentrated only in the capital, Algiers, and other major northern cities, to a health system that works to improve coverage and quality of healthcare [1]. Algeria government pays special attention to the health sector in terms of mobilized resources or adopted policies, with the aim of ensuring adequate care and its free provision in the public sector [1, 2, 3 4, 5], .

With the emergence of computer technologies, the Algerian government attaches particular interest to the digitalization of the health sector [6, 7, 8]. In this regard, the health sector has benefited from several investments, particularly in Information and Communication Technologies (ICT), which have given rise to the launch of electronic projects, to modernize health structures and activities. The majority of Algerian healthcare organizations have recently adopted Electronic Health Record (EHR) information systems. By converting most paper-based records to digital ones and giving users additional management tools. These systems have improved the overall efficiency of the healthcare delivery system through the dynamic representation of information that allows continuous monitoring of a patient's health status by combining everything from the patient's history and treatments to the results and suggestions for additional diagnostic tests [9,10, 11]. However, current EHR in Algeria, store a large number of medical records independently and in different structures and formats, using their digital medical records [7, 8]. Their use is limited for several reasons: the resistance to change of

users (medical secretaries, doctors, nurses, etc.) as well as the problem of interoperability between patient programs [7].

To provide effective and timely health care, health information must be exchanged effectively and efficiently between different health care settings. In Algeria, limited patient access to health data is a major problem, especially when the patient is treated in a structure other than the one he is used to and medical records (consultations, analyses, surgical procedures, etc.) are not shared with other health care structures. In addition, the health institution's information system cannot communicate and share remotely with another information system in the event of an urgent consultation by other health care providers [7].

Interoperability can be defined [10] as the ability of two or more systems or components to communicate with each other and to use the information exchanged. In order to facilitate system interoperability and meaningful data sharing, the syntax (structure) and semantics (meaning) of the data exchanged must be covered by health information standards [12,13,14].

Many healthcare systems are increasingly adopting Electronic Health Records (EHRs) to enhance processes for both patients and doctors, thereby improving care [15]. Despite the numerous global solutions proposed for EHR interoperability, their implementation and adoption within the Algerian community are still limited [9].

In this paper, we examine how ontologies and web service composition can serve as strategic methods to address interoperability challenges, improve data sharing, and foster collaborative healthcare practices.

The aim of this study is summarized in the following points:

- To give an overview of the situation in the health sector, in Algeria.
- To provide current information on various approaches to address semantic interoperability in EHRs explicitly
- To invest in the capabilities of the Semantic Web to improve semantic interoperability in health information systems.
- To develop a prototype facilitating the exchange and sharing of data and information in the health information system in Algeria

The rest of the paper is organized as follows : In section 2 Algeria's Health Status is presented. Section 3 introduces the main concepts and previous work carried-out on Semantic interoperability in healthcare . Section 4 describes a motivation example for implementation and adoption of interoperability. In Section 5, we present the main components of our framework architecture. Finally, Section 4 concludes the paper with suggestions for future work.

2. Algeria's Health Status

Health is a universal fundamental right that Algeria has recognized in all its constitutions since independence. Algeria's healthcare system offers state-guaranteed free services, encompassing both public sector healthcare, such as University Hospital Centers (UHC), and private sector healthcare, including clinics [3, 6]. However, the medical sector in Algeria presents an untapped opportunity for the implementation of new IT systems [7]. Clinics, healthcare institutions, laboratories, analysis centers, and medical practices require modern computerized management solutions powered by robust software to address the challenges in this field, some of which can be mentioned in the following points:

- The patient's medical history (consultations, tests, surgical interventions, etc.) is not recorded in order to gain an idea of his or her health status or to share it with other health institutions, and this may delay appropriate treatment or, in the worst case, not lead to treatment

- The health institution's information system is unable to communicate and share with another information system remotely.
- The absence of proper recording of health information significantly impacts healthcare planning decisions, including determining where to set up a new service, deciding on the implementation of a national screening program, and evaluating the best value for money in health and social care provision.

To facilitate information exchange and the sharing of medical knowledge, hospitals and healthcare providers have aimed to integrate the functions and data of their systems. However, this effort has raised concerns, including data security, data transmission, and network limitations. Among these challenges, the interoperability of systems and data stands out as the most significant barrier to successfully integrating the functions and data of different systems [15].

3. Background and Related Works

Semantic interoperability is the ability to share data between systems and ensure understanding at the concept's level of the domain [16]. It aims to facilitate the sharing of data between systems or organizations and ensure that it is understood and interpreted by all parties involved, using domain concepts, contextual knowledge, and formal data representation [15].

Semantic interoperability in healthcare, is a topic of interest for the eHealth community; Data can be shared, used, and processed from any location while maintaining its original meaning across various systems, apps, and devices. It can be viewed from three different levels [17, 18]: Foundational, Semantic and Structural levels. The prerequisites for securely connecting various systems and exchanging data are described at the Foundational level. While the semantic level Process interoperability allows for the use of publicly defined terminologies, vocabularies, standardized values, and the ability to fully interpret the understanding. The structural level defines the format, syntax, and data that needs to be interpreted at the field level.

Many healthcare systems are progressively using Electronic Health Records (EHRs) to streamline procedures for patients and doctors while facilitating care. The EHR can be considered the natural evolution of the physical medical record, which consisted of paper medical documents for a patient. It consists of several basic elements [19], which include demographic data, medical history, prescriptions, laboratory results, allergies, vital signs, radiology reports, problem lists, and progress notes.

The electronic health record (EHR) can function more efficiently if the syntax and semantics of the data are implemented in an interoperable manner; By ensuring a common understanding of the medical data being managed, it reduces the risk of medical errors. It also efficiently and quickly organizes the collaboration of different healthcare providers (hospitals, government agencies, clinics, general practitioners, etc.) [20]. However, to achieve an electronic medical record, a number of obstacles must be overcome [21], such as understanding a wide range of terminology, resolving ambiguities, and defining and updating concepts. In [22], the authors discuss the interoperability of electronic health record from a management and business standpoint, explaining the links between healthcare interoperability and other areas such as telemedicine, big data, and strategic surveillance.

To accomplish good semantic interoperability between EHRs by establishing a shared vocabulary and conceptual understanding of clinical data, three key components are required: Terminologies, ontologies, and EHR standards [23].

3.1 EHR standards

Currently, there are global initiatives aimed at improving the semantic development of EHR data and, in turn, increasing patient data availability [24]. The global strategy on Digital Health 2020–2025 of

World Health Organization (WHO) and the European Health Data Space (EHDS) regulation proposal is interoperability between health information systems.

Several international standards development organizations (SDOs) such as International Standards Organization (ISO), Comité Européen de Normalisation (CEN) and Health Level Seven (HL7)) have created interoperability standards, which have been widely adopted globally, in order to facilitate the exchange of health information.

3.2 Medical terminology

Medical terminology is a collection of related concepts and words that provide structured medical expressions and a controlled vocabulary to support clinical care, decision-making, search engine optimization, and quality improvement [25]. Terminologies may include clinical reference terminologies or interface terminologies. Clinical reference terminologies is a collection of terms, concepts, and relationships, in human language that have a semantic hierarchy and serve as a common reference point for comparing and aggregating medical data. Examples of reference terminologies: ICD-9-CM, LOINC, SNOMED CT. Interface terminology, also referred to as input terminology, application terminology, or colloquial terminology, is a systematic set of healthcare-related terms that aid clinicians in entering patient data into computer systems. Examples of interface terminology: NANDA, local expressions and preferences.

3.3 Ontology

In the context of knowledge representation and information science, ontology is the formal and explicit specification of a common conceptualization within a particular field. It is basically a structured framework that outlines the entities, concepts, and relationships pertinent to a given field of study. By capturing and representing the underlying semantics of information, ontologies help people and systems communicate by establishing a common vocabulary and understanding [26, 27]. As a central component of the Semantic Web, ontology can be created using any computer description language, including the Web Ontology Language (OWL) and the Resource Description Framework (RDF).

Since ontologies provide consistent semantics and explicit formal models, several works expressed interest in creating distributed and interoperable electronic health record environments based on ontologies. [19] proposed a semantic ontology-based model that can unify different health record environments (EHRs) data formats. [28] developed a method of semi-automated integration to convert different classical data sources (like RDB, XML, and UML) into local ontologies (OWL2). In order to identify related concepts and avoid their duplication in the combined output, these ontologies were then integrated into a global ontological model using syntactic, structural, and semantic similarity measurement techniques. A two-step fusion of electronic medical record data with ontologies is proposed by [29] In the first step, specific data and keywords from various EHR formats are converted into a more easily parsable format (JSON format). In the second step, the obtained JSON data is combined with specialized medical ontologies. Several medical data standards for heterogeneous EHRs were used to implement the multi-step semantic architecture proposed by Kiourtis [30]. The suggested framework included a mechanism for removing explicit space data from ordered EHR datasets and converting them to Common Health Language via ontologies. Furthermore, the use of ontologies has shown promise in mapping scenarios for data access rules. The authors in [31] employ an ontology-based mediation service to enhance the physician's query and utilizes the MiniCon algorithm for the creation of a composite web service. In [32], an agent coordination infrastructure

maps community organizations' access rules to the data using an OWL (Web Ontology Language) ontology.

4. Motivation Example

We present a scenario in which a patient consults at the medical-surgical emergency department of hospital A. He has a history of previous hospitalizations at hospital B, where his entire medical record is kept. This record contains his surgical reports, his treatment protocols, as well as the additional tests he has undergone, including imaging tests and biological evaluations. The patient does not have his medical record at the time of admission, and given the severity of his current health condition and the importance of the medical record to guide the diagnosis, it is essential to obtain information on his medical history and the treatments he has received. Communication between the two hospitals is crucial for the well-being of the patient.

5. The Proposed Model

In our proposed model, hospitals are linked through a semantic interoperability framework built on semantic web services composition, with each patient assigned a unique identifier. It is based on a set of ontologies of domains and services.

Web services are essential to contemporary software architecture, enabling communication and data exchange between various applications and systems via the internet. They promote interoperability, allowing diverse software applications to interact smoothly, regardless of their underlying technologies or platforms. The description of web services can be categorized into two approaches [33]: the first utilizes semantic description languages derived from the semantic web, while the second relies on annotations of web service descriptions that leverage the extensibility features of currently available languages. In our model, we have chosen the approach that utilizes annotations for web service descriptions. Every functionality of a Health Information System (HIS) application is carried out via a web service that is semantically annotated with the local ontology created by the application. Consider that all hospitals have the same web services. Here is some web services used in the HIS:

Table1. Web services for HIS model

Services	Functionalities
Ser_patient	allows for the exchange of patient data
Ser_exammination	allows you to have the different examinations of a patient
Ser_consultation	allows to know the different consultations of a patient
Ser_history	allows you to send back information about a patient's medical history
Ser_result	allows to exchange and know all the results of a patient

A Web service allows users to execute queries. When it comes to executing complex queries that require the integration of multiple web services, the composition of these services becomes essential. The operation of searching for services to compose them is called Web service discovery [34]

The composition of Web services involves creating new web services by integrating two or more existing and published services to enhance their value [35]. Web service composition is classified into two categories [36]: static and dynamic. In static web service composition, the requester must first specify the individual web services involved in the composition, along with their execution order. In contrast, dynamic composition involves determining which web services to combine gradually during the execution of the user's request. In our work, we are interested in dynamic composition of web services. Several approaches for dynamic composition of web services have been presented in [37].

Our proposed interoperability model supports the composition of web services for complex queries by using a mediator based on a global ontology and semantic web services [38]. The execution of a complex query is illustrated in Fig. 1. The different steps involved in the execution are as follows

- When a healthcare staff searches for information about a patient, he/she expresses his/her need in natural language in a query format.
- The query is executed and will be semantically enriched with terms based on our global ontology.
- The web services required to execute the rich query are discovered.
- The discovered web services are composed and executed.
- The required data is then sent to the healthcare staff.

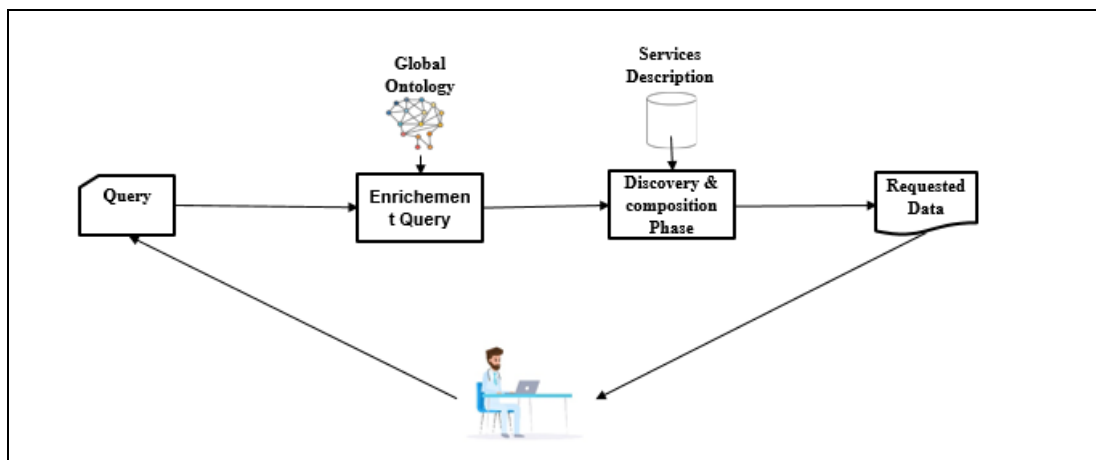


Figure 1. Architecture of the proposed model for HIS interoperability

- Global Ontology

The proposed model uses a mediator based on a global ontology [38]. The global ontology is obtained from local ontologies generated from HIS applications.

After a long discussion with Chirurgical Doctor from the CHU of Batna, Agleria. Where she mentioned all the terms used between hospitals, doctors and all the medical staff on general. In addition, merging the local ontologies of the HIS. The local ontologies are created from the HIS databases. We create a global ontology for our HIS model.

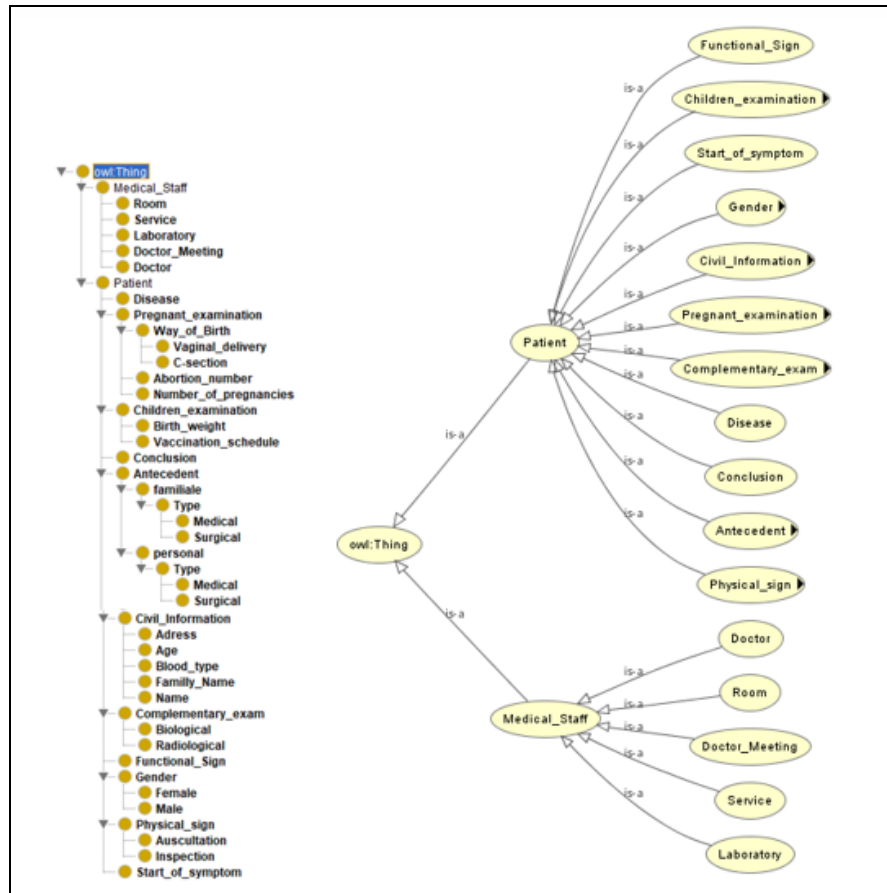


Figure 2. The global ontology for HIS model

5. Conclusion

Health is a universal fundamental right that Algeria has included in all its constitutions since its independence. Through this work, we have clarified the health situation in Algeria. We have highlighted the need to find a solution to improve semantic interoperability between health information systems in Algeria.

Due to the complexity of the interoperability of all healthcare information systems, we have taken in this work, the business process concerning the exchange of data between two hospitals. We have sketched a scenario in which the hospital information systems of two hospitals, exchange data via web services in response to a query from the healthcare staff. A Web service is a software component that allows users to execute queries. The composition of Web services is necessary in the case of execution of complex queries requiring the integration of several web services.

Our proposed interoperability model supports the composition of web services for complex queries by using a mediator based on a global ontology and semantic web services. After extensive consultation with surgical specialists at the CHU of Batna, Algeria, this global ontology was developed. A web service that is semantically annotated with the local ontology derived from each HIS application provides the functionalities of that application. The suggested model encourages communication and information exchange between numerous interconnected HIS components and is made to be extensible.

References

- [1] T. Meziane, *Système de santé en Algérie*, polycopié édité, Dar Konouz Lilkitab pour la Production et Distribution Tlemcen Algérie, ISBN : 978-9931-706-92-02, 2019.
- [2] K. BENMESSAOUD and F.Z. BENCHIKHA, *Les systèmes d'information hospitaliers en Algérie entre mythes et réalités*, Les ouvrages du CRASC, 2022, p.165-185.
- [3] Y. Abbou and B. Brahamia, *Le système de santé algérien entre gratuité des soins et maîtrise des dépenses de santé*. *Insaniyat*, (75-76), 149-171. 2017.
- [4] A. Hamadi, et al. , *Financement de la sante, dépenses de santé et sécurité sociale en Algérie durant la période 2002–2018*, *Journal of Economics and Applied Statistics*,17(3),128, 2020.
- [5] M. Messaili and N. Kaïd Tlilane, *Dépenses publiques de santé et santé de la population en Algérie : une analyse économétrique*. *Santé Publique*,29(3),383 2017.
- [6] A. Zidane, et al. , *Algerian experience in digitizing the health sector*, Article in *المجلة الدولية للبحوث الأكاديمية* · ISSN 2588-1736 EISSN 2661-7269 , pp. 150-172 , June 2023.
- [7] M. SALMI, et al. *Le système de santé en Algérie face au développement du numérique : le dossier du patient informatisé .Cas du CHU de Tizi-Ouzou*. *Revue Française d'Economie et de Gestion* ISSN:2728-0128 Volume3:Numéro5, 2022.
- [8] M. A. Brahmi and H. M. DERGA, *Towards the digitalization of the health sector: Case of the SIHATIC project in the context of the covid 19 pandemic* , *Review MECAS* V°18 / N° 1/ June 2022 .
- [9] M. Aouachria et al. , *The Base Frame of the Algerian E-Health System: Methodology to Select the IHE Integration Profiles and the Standards to Control Interoperability*, *dHealth 2020 – Biomedical Informatics for Health and Care*, 153-160 .
- [10] Sandra V. B. Jardim , *The Electronic Health Record and its Contribution to Healthcare Information Systems Interoperability*, *Procedia Technology* 9 (2013) 940 – 948.
- [11] A. M. Uslu, J. Stausberg. *Value of the Electronic Patient Record: An Analysis of the Literature*. *Journal of Biomedical Informatics*, vol. 41, pp. 675-682, 2008.
- [12] International Telecommunication Union, (2012), *E-Health Standards and Interoperability*. ITU-T Technology Watch Report.2012 [Online]. Available from: www.itu.int/en/ITU-T/techwatch/Pages/ehealth-standards.aspx. Accessed on: 15 September 2024
- [13] R. Ait Abdelouahid et al., *Literature Review: Clinical Data Interoperability Models*, *Information* 2023, 14, 364. <https://doi.org/10.3390/info14070364>.
- [14] *Introduction to Semantic Interoperability*, KNOWLEDGE TOOLS, Pan American Health Organization, 2021, <https://iris.paho.org/handle/10665.2/55417>, Accessed on: 10 September 2024
- [15] Benson T, Grieve G. *Why Interoperability Is Hard*. In: *Principles of Health Interoperability*, 4th ed. Springer, Cham. 2021;21–40.
- [16] ISO 18308:2011(en) *Health informatics — Requirements for an electronic health record architecture*. <https://www.iso.org/obp/ui/#iso:std:iso:18308:ed-1:v1:en>. Accessed September 02, 2024.
- [17] Healthcare Information and Management Systems Society. *What is Interoperability?*. [https:// www.himss.org/resou rces/ inter opera bility- healt hcare](https://www.himss.org/resou rces/ inter opera bility- healt hcare). Accessed September 02, 2024.
- [18] AR. Lee,et al., *Developing a Transnational Health Record Framework with Level-Specific Interoperability Guidelines Based on a Related Literature Review*. *Healthcare (Basel)*. Accessed September 02, 2024. doi: 10.3390/healthcare9010067. PMID: 33450811; PMCID: PMC7828296.
- [19] E. Adel et al.: *Semantic Architecture for Interoperability in Distributed Healthcare Systems*, *IEEE Access* Volume 10, 2022.
- [20] S. Bhartiya, D.Mehrotra, *Challenges and recommendations to healthcare data exchange in an interoperable environment*. *Electron. J. Health Inform.* 2014, 8, 16.
- [21] X . Gansel, et al., *Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review*. *Eur J Clin Microbiol Infect Dis*. 2019;38(6):1023–34.

- [22]Salomi MJA, Claro PB. Adopting Healthcare Information Exchange among Organizations, Regions, and Hospital Systems toward Quality, Sustainability, and Effectiveness. *Technol Invest*. 2020. https://www.scirp.org/html/3-9900884_101997.htm Accessed.
- [23]J. M.do Espírito Santo and C. B. Medeiros, Semantic Interoperability of Clinical Data. In M. Da Silveira, C. Pruski, & R. Schneider (Eds.), *Data Integration in the Life Sciences* (Vol. 10649, pp. 29–37). Springer International Publishing 2017. https://doi.org/10.1007/978-3-319-69751-2_4
- [24]F. A..Fernandes, et al., The role of governance in the digital transformation of healthcare: Results of a survey in the WHO Europe Region, *International Journal of Medical Informatics*, Volume 189, 2024, 105510,ISSN1386-5056,<https://doi.org/10.1016/j.ijmedinf.2024.105510>. (<https://www.sciencedirect.com/science/article/pii/S138650562400173>).
- [25]Bernaldo De Quiros et al, Terminology Services: Standard Terminologies to Control Health Vocabulary, *IMIA Yearbook of Medical Informatics*, 2018:227-33. <http://dx.doi.org/10.1055/s-0038-1641200>.
- [26]H. Liyanage et al., Using ontologies to improve semantic interoperability in health data. *Journal of Innovation in Health Informatics*, 2015, 22(2), 309–315. <https://doi.org/10.14236/jhi.v22i2.159>.
- [27]M.A Haendel, et al., Classification,ontology, and precision medicine. *N.Engl. J.Med.* 2018,379,1452–1462.
- [28]O. El Hajjamy, et al., Integration of heterogeneous classical data sources in an ontological database. In *International Conference on Big Data, Cloud and Applications*; Springer: Cham, Switzerland, 2018; pp. 417–432.
- [29]S.Calcagno et al. , Merging Ontologies and Data from Electronic Health Records, *Future Internet* 2024, 16, 62. <https://doi.org/10.3390/fi16020062>.
- [30]Kiourtis, A.; Mavrogiorgou, A.; Kyriazis, D. Aggregating heterogeneous health data through an ontological common health language. In *Proceedings of the 2017 10th International Conference on Developments in eSystems Engineering (DeSE)*, Paris, France, 14–16 June 2017; pp. 175–181.
- [31]Lydie Simone KONE/TAPSOBA et all, Interoperability approach for hospital information systems based on the composition of web services, *Procedia Computer Science* 219 (2023) 1161–1168.
- [32]V. Urov et al. An Agent Coordination Framework for IHE based Cross-Community Health Record Exchange. In: Moreno A, Isern D, editors. *International conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Valencia: Spain; 2012. p. 29–39.J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [33]K. Mecheri et al., Context-based interoperability of semantic web services, *International Journal of Metadata, Semantics and Ontologies*, 2019.
- [34]J. M. García et al. Improving semantic Web services discovery using SPARQL-based repository filtering. In: *Web Seman. Sci. Serv. and Agents on the World Wide Web*, 17, pp. 12–24, 2012
- [35]Q.Z. Sheng et al., Web services composition: A decade’s overview, *Information Sciences* 280 (2014) 218–238.
- [36]M. Gnana Sagaya Sharmila and S. Justin Samuel, A Trend Analysis on Web Service Composition, *International Journal of Innovative Research in Science, Engineering and Technology*, Volume 3, Special Issue 1, February 2014.
- [37]A. L. Lemos, F. Daniel, and B. Benatallah. Web service composition : a survey of techniques and tools. *ACM Computing Surveys (CSUR)*, 48(3) :33, 2016.
- [38]Khan WA, et al., An adaptive semantic based mediation system for data interoperability among Health Information Systems, *J Med Syst.* 38(8) (2014), 28.

SESSION 2

Environmental Forecasting and Risk Modelling



Interval Markov chain modeling for weather forecasting

Salah Eddine SEMATI¹, Abdelkader GASMI²

¹ salaheddin.semati@univ-msila.dz Address; Laboratory of Signals Analysis and Systems,
Faculty of Mathematics and Informatics, University of M'sila, BO box 166 Ichebilia 28000
Msila, Algeria

² abdelkader.gasmi@univ-msila.dz Address; Laboratory of Pure and Applied Mathematics, Faculty
of Mathematics and Informatics, University of M'sila, BO box 166 Ichebilia 28000 Msila, Algeria

Abstract

This article explores the use of Interval Markov Chains (IMCs) to improve the accuracy and reliability of weather forecasting. IMCs enhance traditional Markov models by incorporating temporal intervals to represent transitions between different weather states (e.g., sunny, rainy, cloudy). This approach better captures the dynamic and uncertain nature of atmospheric processes, allowing for more realistic modeling of weather patterns over time. The study demonstrates that IMCs improve forecasting accuracy by handling uncertainty and non-linear transitions, and can adapt to real-time data, offering more flexible and responsive predictions. The research highlights the potential of IMCs to advance weather prediction models and contribute to more sophisticated meteorological tools.

Keywords: Weather forecasting, State, Markov Chain, IMC, Prediction.

1. Introduction

In the realm of weather forecasting, the accurate prediction of atmospheric conditions stands as a critical imperative for applications spanning agriculture, disaster management, and daily life. However, the inherent intricacies and uncertainties associated with atmospheric processes present formidable challenges to meteorologists seeking to improve the precision of their forecasts. Recent advancements in modeling techniques have spurred exploration beyond conventional methods, with Interval Markov Chains (IMCs) emerging as a promising avenue for addressing the limitations of traditional approaches. Interval Markov Chains offer a sophisticated and probabilistic framework that introduces a temporal dimension to weather pattern evolution. Unlike conventional Markov Chain models, which assume instantaneous and deterministic state transitions, IMCs consider intervals, allowing for a more realistic representation of the gradual and uncertain nature of atmospheric changes [1]. This temporal nuance enables the modeling of complex dependencies inherent in meteorological data,

thereby enhancing the capacity to capture the dynamic behavior of weather systems. This article embarks on an exploration of the application of Interval Markov Chains as a powerful tool to elevate the accuracy and reliability of weather predictions. The decision to integrate IMCs into forecasting methodologies stems from their unique ability to accommodate the temporal complexities of atmospheric transitions. As we delve into the subsequent sections, the methodology of incorporating Interval Markov Chains in weather forecasting will be elucidated, supported by empirical evidence that highlights the efficacy of this approach when compared to traditional models [2]. This endeavor contributes to the ongoing dialogue in meteorological research, striving to refine our understanding of atmospheric dynamics and bolster our predictive capabilities. By adopting Interval Markov Chains, we aim to pave the way for more nuanced and adaptive weather forecasting models that can better align with the unpredictable nature of weather phenomena, thus advancing the field and its applications across diverse sector

2. Background

Markov chain [2] was invented by A.A. Markov; a Russian Mathematician, in the early 1900's to predict the behavior of a system that moves from one state to another by considering only the current one. This chain is a mathematical model that describes a system that transitions between different states over time. It is a type of stochastic process that is characterized by the Markov property. The probabilities are usually represented by a transition matrix (we will talk about it in the next title). This model have a wide range of applications in various fields, including physics, biology, finance, and computer science. They are often used to model random processes, such as the behavior of molecules in a chemical reaction, the spread of a disease in a population, or the performance of a stock market over time. They are also used in natural language processing, where they can be used to model the probability of a sequence of words occurring in a sentence, and in machine learning, where they can be used to generate realistic texts or to perform image segmentation.

2.1 Transition matrix

A transition matrix in a Markov chain is a square matrix that represents the probabilities of transitioning between states in the chain. The matrix is usually denoted by the symbol P and has dimensions $n \times n$, where n is the number of states in the Markov chain. Each element $P_{i,j}$ of the transition matrix represents the probability of transitioning from state i to state j in one-step, the sum of the probabilities of transitioning from state i to all other states must equal one, since the system must be in one of the states at any given time. For example, suppose we have a Markov chain with three states A, B, C, and the transition matrix P is given by:

$$P = \begin{pmatrix} P_1^1 & P_1^2 & P_1^3 \\ P_2^1 & P_2^2 & P_2^3 \\ P_3^1 & P_3^2 & P_3^3 \end{pmatrix}$$

The element in the first row and second column P_{12}^2 represents the probability of transitioning from state A to state B in one-step. Similarly, P_{23}^3 represents the probability of transitioning from state B to state C in one-step. This matrix is a fundamental concept in Markov chain analysis, as it allows us to calculate the probabilities of various events, such as the probability of reaching a particular state after a certain number of steps, or the probability of the system being in a certain state at a given time

2.2 The interval Markov chain

Interval Markov chains (IMCs) are a type of mathematical model used to analyze systems that evolve over time and undergo discrete changes or transitions. These systems are characterized by the presence of uncertainty and randomness in their state transitions. IMCs are particularly useful in modeling systems where transition probabilities may vary within certain intervals, rather than having fixed values.

2.2.1 Definition 1

Following de Campos et al [3], the probability interval is defined by:

$$P = [\underline{P}_i^j, \overline{P}_i^j]$$

In which $\underline{P}_i^j$ and $\overline{P}_i^j$ verifying:

$$\begin{aligned} 0 \leq \underline{P}_i^j \leq \overline{P}_i^j \leq 1, \forall 1 \leq i, j \leq 3 \\ \left\{ \sum_i \underline{P}_i^j \leq 1 \leq \sum_i \overline{P}_i^j \leq 1, \forall 1 \leq j \leq 3 \right\} \end{aligned}$$

The transition probability of any system from an actual state i to another future state j by one-step is noted by P_{ij}^j , in which $\underline{P}_i^j \leq P_{ij}^j \leq \overline{P}_i^j$, and then the one-step transition interval probability matrix can be expressed as:

$$P = \begin{pmatrix} [\underline{P}_1^1, \overline{P}_1^1] & \dots & [\underline{P}_1^n, \overline{P}_1^n] \\ [\underline{P}_n^1, \overline{P}_n^1] & \dots & [\underline{P}_n^n, \overline{P}_n^n] \end{pmatrix}$$

2.2.2 Definition 2

Let the initial interval probability distribution is:

$\pi(0) = ([\underline{P}_1^0, \overline{P}_1^0], \dots, [\underline{P}_n^0, \overline{P}_n^0])$, one-step transition probability interval matrix is p , the next period distribution could be predicted as:

$$\pi(1) = \pi(0) \times P$$

The distribution after k period should be expressed as:

$$\pi(k) = \pi(k-1) \times p$$

We can deduce from the above formula that the weather forecasting of any period in the future can be predicted if the one-step transition probability matrix P and the initial weather condition $\pi(0)$ are known. And, if the one-

step transition probability matrix is known, the weather forecasting after a certain period can be calculated using the formula.

$$\begin{cases} \pi = \pi \times P, \pi = ([P_1, \overline{P_1}], \dots, [P_n, \overline{P_n}]) \\ \sum_i^n P_i \leq 1 \leq \sum_i^n \overline{P_i} \end{cases}$$

2.3 Weather forecasting

Weather forecasting plays a pivotal role in our daily lives, impacting sectors ranging from agriculture to transportation. The accuracy and reliability of weather predictions have significantly improved over the years, thanks to ongoing research and technological breakthroughs. The history of weather forecasting dates back to ancient civilizations, where rudimentary methods were employed to predict atmospheric conditions. Ancient civilizations used various methods to predict the weather, although their approaches were often based on observations and patterns rather than the sophisticated technology we have today. Here are some common methods used by ancient cultures:

1. Observation of Natural Signs:

- (a) **Cloud Patterns:** People observed the appearance, movement, and types of clouds. Certain cloud formations were associated with specific weather patterns.
- (b) **Animal Behavior:** Changes in animal behavior were often considered indicators of impending weather changes. For example, birds flying low or animals seeking shelter could be signs of an approaching storm.

2. Astronomy:

- (a) **Celestial Bodies:** Some cultures looked to the stars, moon, and sun for clues about upcoming weather. Certain celestial events were thought to be connected to weather changes.

3. Plant and Environmental Indicators:

- (a) **Plant Life:** The growth and behavior of plants were often observed. For instance, the closing of flowers or the behavior of pine cones was considered indicative of changes in humidity or upcoming rain.
- (b) **Wind Direction:** Wind patterns and changes in wind direction were carefully observed as they could signal changes in the weather.

4. Natural Phenomena:

- (a) **Earthquakes:** In some cultures, earthquakes were believed to be linked to changes in weather patterns.
- (b) **Volcanic Activity:** Volcanic eruptions were sometimes associated with specific weather changes.

5. Local Knowledge and Folk Wisdom:

- (a) **Proverbs and Sayings:** Many cultures developed proverbs and sayings based on accumulated wisdom about weather patterns. For example, "Red sky at night, shepherd's delight; red sky in the morning, shepherd's warning."

6. Instrumental Measurements:

- (a) **Simple Instruments:** Some ancient cultures used basic instruments like rain gauges, wind vanes, and thermometers, although these were not as advanced as modern equivalents.

It is important to note that while these methods were based on keen observations, they lacked the scientific

Understanding and technology we have today. Modern meteorology relies on advanced instruments, satellite imagery, and computer models to make accurate weather predictions. Over time, advancements in instrumentation, satellite technology, and computational modeling have revolutionized the field [4]. Weather forecasting relies on a combination of observational data, numerical models, and sophisticated algorithms. Observational data from ground-based stations, satellites, and remote sensing devices form the foundation of weather forecasting. The integration of data from various sources enhances the accuracy of predictions [5].

2.4 Harnessing the Power of Interval Markov Chains for Weather Forecasting

Weather forecasting has evolved significantly over the years, driven by advancements in technology, data assimilation methods, and mathematical modeling techniques [6]. The accurate prediction of atmospheric conditions is crucial for various sectors, including agriculture, transportation, and emergency management. In recent times, researchers have been exploring innovative approaches to enhance the precision and reliability of weather forecasts. In our research, we want to add one of the most promising techniques, using Interval Markov chains.

Traditional weather forecasting relies on numerical weather prediction (NWP) models that simulate the complex interactions between atmospheric components. These models use mathematical equations to represent physical processes, and observations from satellites, weather stations, and other sources are assimilated to initialize the simulations [6]. Recognizing the inherent uncertainty in weather predictions, probabilistic forecasting has gained prominence. Instead of providing a single deterministic forecast, probabilistic approaches offer a range of possible outcomes along with associated probabilities. This shift towards probabilistic forecasting has opened the door for advanced statistical techniques like Interval Markov Chains [7]. Interval Markov Chains (IMCs) are an extension of traditional Markov Chains, a stochastic model that represents a system transitioning between different states with associated probabilities. IMCs introduce the concept of intervals, allowing for a more nuanced representation of uncertainty. In the context of weather forecasting, IMCs can capture the dynamic and uncertain nature of atmospheric processes, providing a valuable tool for generating probabilistic forecasts [6]. By incorporating interval based uncertainty representations, these models can better capture the variability and unpredictability inherent in atmospheric conditions, especially for extreme events and complex weather phenomena [8]. As technology continues to advance, the potential for IMCs to revolutionize weather forecasting is becoming increasingly evident [9]. The interval probability distribution of the initial state can be expressed as:

$$\pi(0) = \left(\left[\frac{1}{N}, \frac{1}{N} \right], \left[\frac{1}{N}, \frac{1}{N} \right], \left[\frac{1}{N}, \frac{1}{N} \right] \right)$$

3 Utilization of the interval Markov chain model

The following fig.1 shows a sample Web graph represent the weather in Msila, containing three situations namely: rainy, sunny, and cloudy.

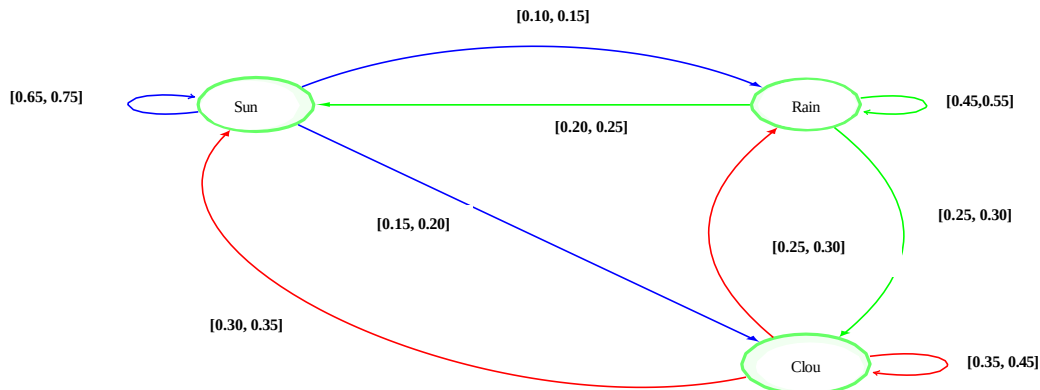


Figure 1: weather in Msila

Let $\bar{P}_{(A,B)}$, be the greater probability that the weather will change from state A to state B, and $\underline{P}_{(A,B)}$, be the smaller probability of the weather changing from state A to state B, and we have:

$$\begin{aligned}\bar{P}_{(sunny,sunny)} &= 0.75, \underline{P}_{(sunny,sunny)} = 0.65 \\ \bar{P}_{(sunny,rainy)} &= 0.15, \underline{P}_{(sunny,rainy)} = 0.10 \\ \bar{P}_{(sunny,cloudy)} &= 0.20, \underline{P}_{(sunny,cloudy)} = 0.15 \\ \bar{P}_{(rainy,rainy)} &= 0.55, \underline{P}_{(rainy,rainy)} = 0.45 \\ \bar{P}_{(rainy,sunny)} &= 0.25, \underline{P}_{(rainy,sunny)} = 0.20 \\ \bar{P}_{(rainy,cloudy)} &= 0.30, \underline{P}_{(rainy,cloudy)} = 0.25 \\ \bar{P}_{(cloudy,cloudy)} &= 0.45, \underline{P}_{(cloudy,cloudy)} = 0.35 \\ \bar{P}_{(cloudy,sunny)} &= 0.35, \underline{P}_{(cloudy,sunny)} = 0.30 \\ \bar{P}_{(cloudy,rainy)} &= 0.30, \underline{P}_{(cloudy,rainy)} = 0.25\end{aligned}$$

3.1 The Algorithm for computing the interval probability vector after m transitions

By using the interval arithmetic operation given in Nirmala et al. [10] and by extending Kaucher[11] interval arithmetic. This arithmetic operation can satisfy group properties with respect to addition and multiplication operations and satisfying the distributive relations between intervals, while maintaining the inclusion monotonicity. We give an iterative algorithm to find the prediction for any given future iteration.

For s=1: m do

For i=1: n do

For j=1: n do

$$\begin{aligned}\pi_i^s &= \pi_i^{s-1} \times \min \left(\pi_i^{k-1} \times \underline{P}_j^i, \pi_i^{k-1} \times \bar{P}_j^i, \pi_j^{k-1} \times \underline{P}_i^i, \pi_j^{k-1} \times \bar{P}_i^i \right) \\ \pi_i^s &= \pi_i^s + \max \left(\pi_i^{k-1} \times \underline{P}_j^i, \pi_i^{k-1} \times \bar{P}_j^i, \pi_j^{k-1} \times \underline{P}_i^i, \pi_j^{k-1} \times \bar{P}_i^i \right)\end{aligned}$$

end for end for end for

3.2 An Illustrative example

We are now going to apply all of this to our example of three state.

Table 1: Estimation probability

state	Sunny	Rainy	cloudy
first Iteration	$\left[\frac{1}{3}, \frac{1}{3}\right]$	$\left[\frac{1}{3}, \frac{1}{3}\right]$	$\left[\frac{1}{3}, \frac{1}{3}\right]$
to Sunny	[0.65, 0.75]	[0.20, 0.25]	[0.30, 0.35]
to Rainy	[0.10, 0.15]	[0.45, 0.55]	[0.25, 0.30]
to Cloudy	[0.15, 0.20]	[0.25, 0.30]	[0.35, 0.45]
second Iteration	[0.4833, 0.5833]	[0.2000, 0.2500]	[0.2167, 0.2667]

Then the interval weather forecasting after k steps can be presented by the formula:

$$\pi^k = \pi^0 \times P^{k-1}$$

$$\pi^k = \left(\begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix} \right) \times \begin{pmatrix} [0.65, 0.75] & [0.10, 0.15] & [0.15, 0.20] \\ [0.45, 0.55] & [0.20, 0.25] & [0.25, 0.30] \\ [0.35, 0.45] & [0.30, 0.35] & [0.25, 0.30] \end{pmatrix}$$

After k period, the interval weather forecasting can be calculated by applying the proposed algorithm. Some obtained results for various periods are illustrated in Table2

Table 2: Some results for deferent iterations

Iteration K	Interval probability distribution π^k
1	([0.4833,0.5833],[0.2000,0.2500],[0.2167 ,0.2 667])
2	([0.4800,0.6950],[0.1883, 0.1983],[0.2167,0.2217])
4	([0.4595,0.7526],[0.1772,0.1880],[0.2100,0.2126])
6	([0.4461,0.7713],[0.1776,0.1848],[0.2058,0.2145])
8	([0.4390,0.7803],[0.1790,0.1824],[0.2032,0.2162])
12	([0.4340,0.7866],[0.1799,0.1807],[0.2014,0.2174])
16	([0.4329,0.7880],[0.1802,0.1803],[0.2009,0.2177])
20	([0.4326,0.7884],[0.1802,0.1802],[0.2008,0.2177])

We can give each situation its final value after the state of stability according to the following table:

Table 3: interval weather forecasting after the stability state

state	Sunny	Rainy	cloudy
The interval weather forecasting	[0.4325,0.7885]	[0.1802,0.1802]	[0.2008,0.2178]

4 Conclusion

In conclusion, this research has delved into the promising realm of utilizing interval Markov chains for weather forecasting, combining principles from computer science and mathematics to address the inherent uncertainties associated with atmospheric dynamics. The proposed model not only embraces the stochastic nature of weather phenomena but also introduces the concept of intervals, allowing for a more robust representation of prediction uncertainties. While this study marks a significant step forward, there are still avenues for further exploration and refinement. Future work should focus on expanding the dataset, refining the model parameters, and exploring novel techniques for enhancing the computational efficiency of the proposed interval Markov chain approach. In conclusion, the findings presented herein underscore the potential of interval Markov chains as a valuable tool for advancing the accuracy and reliability of weather forecasting. As we continue to bridge the gap between computer science and meteorology, the collaborative efforts of researchers in both fields will undoubtedly contribute to the evolution of more sophisticated and robust forecasting models, ultimately benefiting society through improved preparedness and resilience to the impacts of changing weather patterns.

References

- [1] Johnson, R. Brown, Comparative Analysis of Interval Markov Chains and Traditional Models in Weather Forecasting. *Weather and Climate Journal*, Volume (Issue), Page Range.
- [2] Smith. J. A Historical Overview of Weather Forecasting Methods. *Journal of Meteorological Research*, 35(2), 145-162
- [3] R. Norris. *Markov Chains*. Cambridge University Press. 1–4, 1996.
- [4] Campos, Luis M, Huete, Juan F, Moral, Serafín. Probability intervals: a tool for uncertain reasoning. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 2(2).167–196 (1994).
<https://doi.org/10.1142/S0218488594000146>
- [5] Johnson. M, Advancements in Satellite Technology for Weather Observations. *International Journal of Remote Sensing*, 42(8), 1785- 1802(2020)
- [6] Wilks, D. Statistical Methods in the Atmospheric Sciences. *Academic Press* (2011).
- [7] Joliffe, I. T, Stephenson, D. Forecast Verification: A Practitioner's Guide in Atmospheric Science. *John Wiley and Sons* (2012).
- [8] Roulston, M. S, Smith, L. A, estimating the quality of ensemble forecasts. *Monthly Weather Review*, 130(6), 1653-1660(2002)
- [9] Ferro, C. A. Stephenson, D. B, Extremal dependence indices: Improved verification measures for deterministic forecasts of rare events. *Weather and Forecasting*, 26(5), 699-713.(2011)
- [10] T. Nirmala, D. Datta, H. Kushwaha, K. Ganesan, on the solution of system of interval linear equations. *Int. J. Appl. Eng. Res.* 10(6), 15777–15797(2015)
- [11] Kaucher. *Interval analysis in the extended interval space IR*. In: *Fundamentals of numerical computation (computer-oriented numerical analysis)*. Pp.33–49(1980).
<https://doi.org/10.1007/978-3-7091-8577-3-3>.

Fuzzy logic approach for forest fire risk estimation in Algeria

Harbouche Khadidja^{1✉}, Khentout Chabane¹, Haddad Dhai Eddine¹

¹ University Setif 1 Ferhat Abbas UFAS1, Computer science Department, Setif, Algeria
khadidja.harbouche@univ-setif.dz

Abstract

Every year, Algeria frequently experiences natural disasters that result in significant loss of life. Among these disasters, forest fires have significantly contributed to human and economic losses, as well as to the degradation of the ecosystem. Effective disaster management for forest fires is essential to reduce damage both before and after their occurrence. Mitigation is a crucial phase in the forest fire management process. During this phase, we can estimate the risk of a given region affected by this disaster. Our contribution here is to present a model based on fuzzy logic to estimate the risk of wildfires in Algeria. This system uses six parameters that are key indicators for fire susceptibility: altitude, slope, exposure, fuel (vegetation), human presence expressed by distance to settlements, and roads. Our system has been tested in the Setif region, despite it being characterized by a significant lack of data.

Keywords: Forest fire risk, Fuzzy logic, Susceptibility parameters, Dagorne model, Erten model

1. Introduction

Fire is one of the most devastating factors contributing to forest degradation in Algeria. This phenomenon has been observed for several decades.

The biodiversity of plant formations in Algeria (more particularly northern Algeria) is very rich, which represents a very important biological and genetic heritage. In addition, we cannot ignore the increased importance of the forest in its social dimension, particularly during tourist influxes in the summer season. The protection of this forest bio-heritage therefore constitutes a real economic, ecological and social challenge [1]. Several models make it possible to estimate the risk of forest fires. Among the most used are the Dagorne and the Erten models. Our contribution, through this work, concerns a quantitative and qualitative estimation model integrating objective parameters: fuel (vegetation), humans (settlements and roads) and topography (altitude, slope, exposure). This model is built using fuzzy logic for a better formalization of human reasoning.

The rest of the paper is organized as follows: Section 2 presents the chronology, the causes and the risk of forest fires in Algeria, Section 3 shows some wildfires risk models widely deployed in academic works in Algeria (Dagorne and Erten), Section 4 is devoted to our contribution by describing the different steps of our fuzzy based model and in Section 5 an experimentation of our system on the region of Setif is presented. We close with a conclusion.

2. Forest fires in Algeria

2.1 Chronology of Forest Fires in Algeria

The forest fires are not a recent issue in Algeria, which is one of the few countries with statistics on forest fires spanning over a century. According to [2], several analyses of wildfire statistics conducted by Marc in 1916, Boudy in 1948, Grim in 1989, and Meddour-Sahar in 2008 have allowed us to trace their history over a period of 139 years (from 1876 to 2014), with data gaps for the years 1950 and 1962. Statistical analysis highlights a relative stability in burned areas, as shown in Figure1 (a).

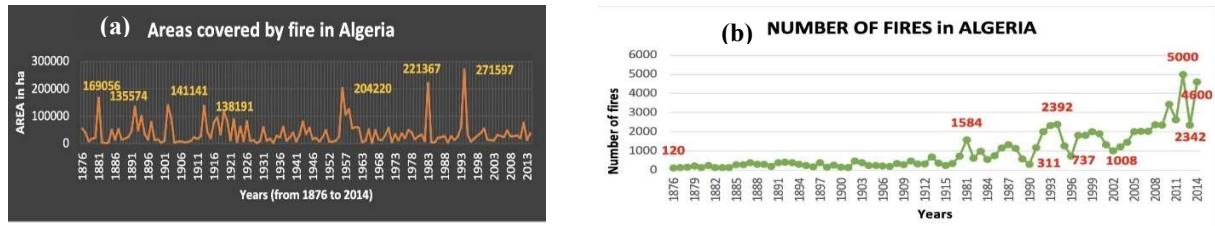


Figure 1. Chronology of forest fires in Algeria

Figure 1 (b) demonstrates a clear trend towards an increase in fire frequency, noted in recent decades, without knowing the exact causes. This will have serious consequences on an already fragile ecosystem if we do not control this ecological factor.

2.2 The wildfires risk in Algeria

In Algeria, the risk of forest fires is spatially concentrated mainly in the coastal northeastern region, from Tizi Ouzou to El Tarf, corresponding to highly wooded and rugged areas, with a high population density. It is the most affected with 50.06% of fires risk, followed by the Central-North region with 28.21%, and finally the Northwest region with 21.73%.

According to [3], [4] and [5] the distribution of the causes of fires in Algeria for the period 1979 to 2020 shows the importance of causes of unknown origin ($\approx 80\%$ in 2020). For known causes (according to official statistics or field surveys), and as shown in Figure 2, we can cite:

- The natural origin of the phenomenon or the climatic conditions.
- Anthropogenic.
- Involuntary causes
- Negligence
- Deliberate fires (malice)

2.3 Favorable factors for forest fires in Algeria

Combustion is a physico-chemical phenomenon. For a fire, especially a forest fire, to start, three ingredients are essential: A fuel (vegetation), an oxidizer (air and the oxygen present in it at 21%) and an activation energy (a heat source “spark” for the reaction to take place). In the event of prolonged drought, on the one hand, there is an atmosphere containing flammable substances, and on the other, very dry plants that are highly flammable. Additionally, the concentrations of oxidizer and fuel must fall between the lower and upper flammability limits.

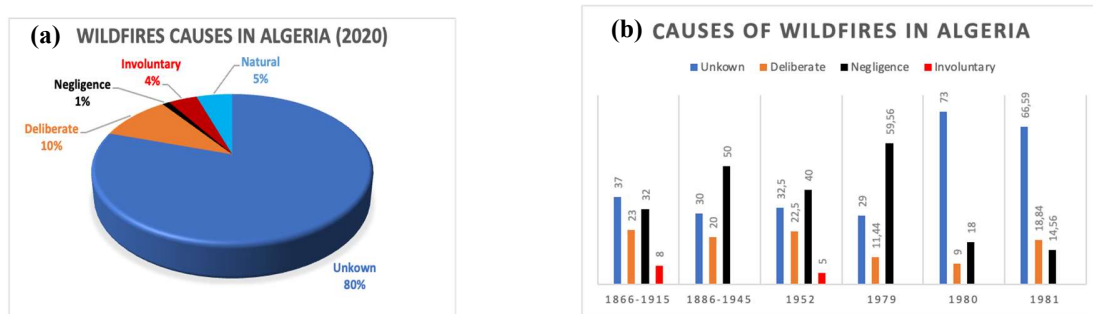


Figure 2. Wildfires causes in Algeria

3. Forest fires risk models

Several studies have been conducted to determine the factors for evaluating the risk of forest fires, particularly in the Mediterranean region. Among the models that have been widely adopted and used in Algeria, the following can be mentioned:

3.1 The Dagorne and Duche Model

This risk index model [7] was proposed and tested on the forested areas of the Mediterranean region. It integrates three indices (as shown in Figure3 (a)): TopoMorphological Index (MI), Human Index (HI) and Combustibility Index (CI). Its risk index (RID) is based on the following formula (Eq. 01):

$$RID = 5 * CI + 2 * HI + MI \quad (1)$$

3.2 The Erten Model

Here, the modeling of forest fire risk is based on several parameters: ecological parameters (flammability and combustibility of vegetation types) related to the type of vegetation ($Tveg$), station parameters (slope, exposure, and topomorphology), and anthropic factors related to the distance from roads (Dr : distance from roads) and distance from dwellings (Dd) [8]. Its risk index (IRE) is based on the following (Eq. 02):

$$IRE = 7 * Tveg + 5 * (P + E) + 3 * (Dr + Da) \quad (2)$$

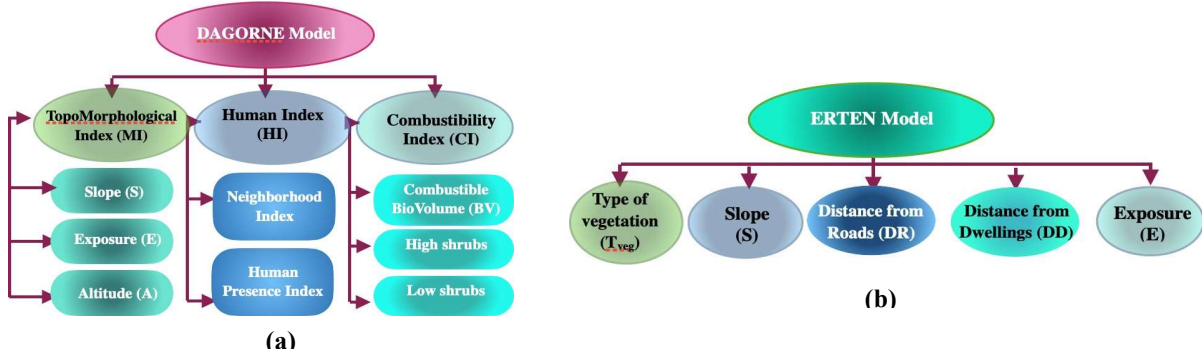


Figure 3. Forest fires risks models

4. Contribution: The fuzzy model

Fuzzy logic is an extension of Boolean logic, introduced in 1965 with the proposal of “fuzzy set theory” by Lotfi Zadeh [9]. It gives very appreciable flexibility to reasoning while taking into account imprecisions and uncertainties.

One of the benefits of using fuzzy logic to formalize human reasoning is that the rules are expressed in natural language. Fuzzy logic is an appropriate way to draw up a wildfire risk estimation according to available factors, namely: topography, fuel (type of vegetation) and human activities.

4.1 Fuzzy risk parameters

As we saw above, there is a close relationship between the index of a forest fire risk with three factors: anthropic factor, the topography of the forest factor and type of vegetation. Therefore, the number of forest fires is proportional to six parameters: slope, altitude, exposure, human neighborhood, human presence and vegetation. These parameters are assigned the same degree of importance in modeling forest fire risk.

4.1.1 Vegetation: Forest fuel consisting of dry leaves, fallen tree branches, and dry wood on the forest floor significantly contributes during the dry season. The ability of plants to generate or propagate fire and to burn through heat transfer depends strongly on their morphological and functional characteristics [10]. We can classify vegetation according to humidity and therefore its flammability and combustibility as

shown in Figure4 (a). The Combustibility Index (CI) is calculated from the moisture content of the vegetation (HV), and the quantity of fuel biomass (BV).

$$CI = 20 * (BV * HV) \quad (3)$$

4.1.2 Slope: The slope conditions the inclination of the flames relative to the ground and thus their propagation speed. The steeper the slope, the higher the risk of fire spread [3]. The slope values expressed as percentage (degree of inclination) are classified as shown in Figure4 (b).

4.1.3 Exposure: In the northern hemisphere, a slope facing south is exposed to the sun, while a north-facing slope is shaded. This influences air temperature and humidity, which in turn affects the risk of fire ignition and spread (Figure4 (c)).

4.1.3 Altitude: Slope gradient usually varies with altitude, but not monotonically. The strength of the relationship varies, and in dissected plateaus, the overall correlation is negative. According to [11], the higher the elevation the lower is the probability of fire. Thus, there is a direct relationship between an increase in forest fire events and an increase in elevation. The altitude is classified into five categories with an interval of 100 m: <800, 800– 900, 900–1000, 1000–1100, and >1100 m.

4.1.4 Human neighborhood: The road networks are intended to allow access to forests (tourism, study ...). They contribute to the destruction of forests following intensive human activities which contribute to the outbreak of fires (voluntary or not). The distance from the road, expressed in meters, is classified to the five classes: <50, 50–100, 100–400, 400–1000, and >1000 m.

4.1.5 Human presence: Forest dwellers (villages) can cause accidental or non-accidental fires especially in hot seasons as a result of cooking, cigar butts; coaling ... Distance from Dwelling is one of the most important human factors that reflect the intensity of human pressure on forest ecosystems. In this regard, The Euclidean distance from the settlement, expressed in meters, is divided into five classes: <50, 50–100, 100–200, 200–500, and >500 m.

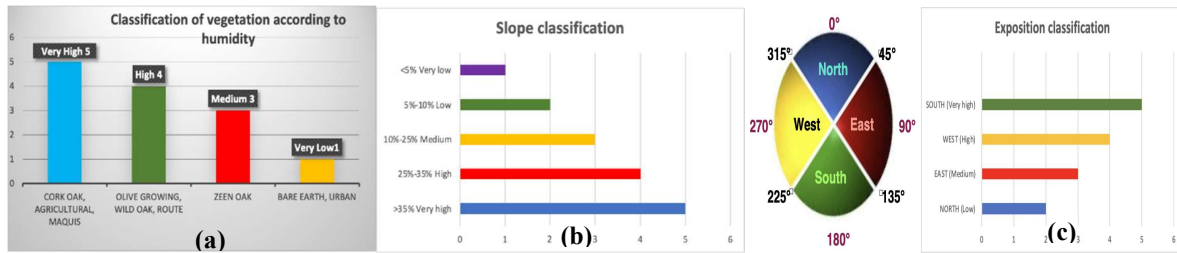


Figure 4. Fuzzy parameters classification

4.2 The fuzzy system

The fuzzy system for forest fire risk estimation consists of four main phases:

4.2.1 The fuzzifier: In this phase, we will convert the crisp input values to fuzzy value.

The inputs to the fuzzy logic system are:

- The Slope S: Figure 5(a) shows the membership function used for this parameter. The membership functions are determined to be “Very Low” (VLs), “Low” (Ls), “Medium” (Ms), “Steep” (S), and “Very Steep” (VS).
- The Altitude (A): that is determined, as shown in Figure5 (b), by the membership functions selected as “Very Low” (VL_a), “Low” (L_a), “Medium” (M_a), “High” (H_a), and “Very High” (VH_a).
- The Exposure (E): Figure5 (c) shows the membership function used for this parameter. The

membership functions are determined to be "North" (N), "East" (E), "West" (W) and "South" (S).

- The Fuel (F): that is determined, as shown in Figure5 (d), by the membership functions selected as "Low" (Lf), "Moderate" (Mo), "Medium" (Me), "High" (Hf) and "Very High" (VHf).
- The Neighborhood (N): that is determined, as shown in Figure5 (e), by the membership functions selected as "Very Close" (VC), "Close" (C), "Medium" (Mn), "Far" (F) and "Very Far" (VF).

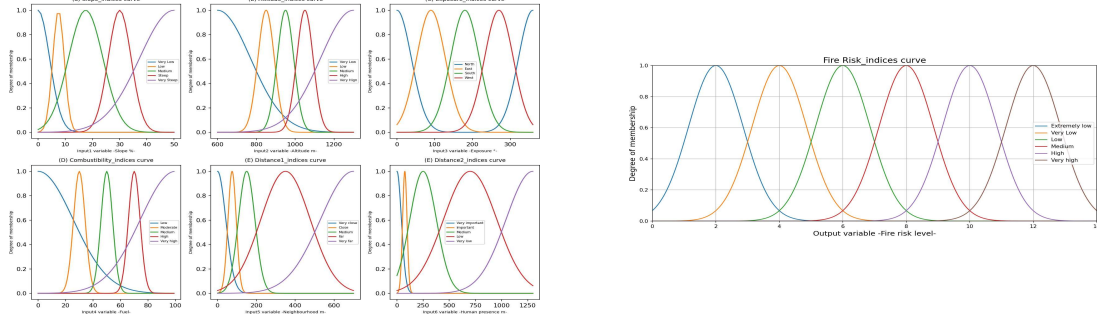


Figure 5. Fuzzy risk system input/Output

- The distance from Dwellings (D): Figure5 (f) shows the membership function used for this parameter. The membership functions are determined to be "Very Important" (VI), "Important" (I), "Medium" (Md), "Low" (Ld) and "Very Low" (VLd).

The input linguistic values are converted into fuzzy input values using the input membership function $f_{variable}(x)$ given in Eq.4.

$$f_{variable}(x) = e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \quad \text{Where, } \mu \in \{2, 4, 6, 8, 10, 12\} \text{ and } \sigma = 0.86 \quad (4)$$

The values of μ and σ for the different membership functions of each variable is given in Table1:

Table 1. Gaussian membership functions parameters

Variable	μ	σ
Slope(S)	0, 7.5, 17.5, 30, 50	4.24, 2.12, 6.36, 4.24, 12.72
Altitude(A)	600, 850, 950, 1050, 1300	169.84, 42.46, 42.46, 42.46, 169.84
Exposition (E)	0 or 360, 90, 180, 270	38.22, 38.22, 38.22, 38.22
Fuel or Vegetation (F)	0, 45, 55, 65, 100	34, 4.25, 4.25, 4.25, 25.5
Distance dwellings(D)	0, 75, 250, 700, 1300	42.47, 21.24, 127.41, 255.82, 255.82
Distance to roads(N)	0, 75, 150, 350, 700	42.47, 21.24, 42.47, 127.41, 169.88

The output of the fuzzy system is the fire risk level (FRL), as we can see in Figure5, it takes six values: Extremely Low (EL), Very Low (VLR), Low (LR), Medium (MR), High (HR) and Very High (VHR). The Gaussian membership function is used as membership function for the output, as shown in Figure5. This function is defined in the universe of discourse $X = [0..14]$ given in Eq.5.

$$F(x) = e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \quad (5)$$

4.2.2 The inference: The fuzzy knowledge base is made up of linguistic variables and fuzzy rules. We use a set of IF-THEN rules, such as:

IF (F is VHf or F is Hf) and (D is VI or D is I) and (N is VC or N is C) and (E is S) and (S is VS or S is S) and (A is VLd or A is Ld) **THEN** FRL is VHR

Based on the fuzzy rules set and input membership functions, fuzzy outputs are obtained.

The choice of the fuzzy operators allows us to determine the inference engine, which is generated, in our case using the Min/Max Zadeh operators.

We apply the fuzzy rules, with the MIN/MAX operators and the different outputs Y1, Y2, Y3, Y4, Y5, and Y6 are obtained.

4.2.3 Defuzzifier: In this phase, the Mean of Maxima (MOM) method will be applied [12] [13]. The defuzzified value is taken as the element with the highest membership values. If there is more than one element having maximum membership values, we can get the mean value of the maxima as follows:

$$Y = \text{MAX}(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) \quad (6)$$

We calculate the inputs (abscissas) named A (Eq. 8) corresponding to the outputs Y given by Eq.(7):

$$Y = e^{(A-\mu)^2/1.48} \quad \text{where, } \mu \in \{2, 4, 6, 8, 10, 12\} \quad (7)$$

$$A = \mu \pm \sqrt{1.48 * \text{Ln}(Y)} \quad (8)$$

4.2.4 The output: The result of the last step of the fuzzy system is to interpret the effect of the topographic, human and vegetation factors on the wildfire risk. There are six effects: Extremely Low, Very Low, Low, Medium, High and Very High. Thus, we obtain the estimation value of the forest fire risk.

5. Experimentation

5.1 Location of the study area

The study area is located in the eastern Algerian highlands. It is one of the fifty-eight administrative regions (wilayas) in Algeria. Setif is geographically located between 35°36'58" to 36°35'45" North latitude and 4°43'52" to 6°1'37" East longitude, where it is bordered by Bordj Bou Arreridj to the west, to the north by the Bejaia and Jijel to the south by Msila and Batna, and to the east by Mila.

The relief of the wilaya of Setif is divided into three main zones: the mountainous zone, the high plains zone and the South and Southeast zone.

5.2 Geographic distribution of national forests

Forest vegetation distribution reflects the climatic conditions and topographic of the region where it is located. The Setif region presents a mixed forest ecosystem of Aleppo pines (*Pinus halepensis*) and holm oak (*Quercus ilex*), it is also made up of Atlas cedar, Zen oak, cork oak and Numidian fir [14].

5.3 Experimental Data

Using OSM [15], we extracted polygons representing forests in our study area using QGIS. Then, a total of 80982 randomly spaced sample points were generated over these polygons. Topographic data (see Figure6) expressed by elevation, slope and aspect were obtained using the NASA Shuttle Radar Topography Mission (2013) Global 30m Digitized Elevation Map (DEM) [16]. To calculate the NDVI, Landsat 8-9 OLI/TIRS C2-L2 satellite imagery courtesy of the U.S. Geological Survey was obtained through the USGS web portal EarthExplorer. Also, to measure the distance from the nearest dwellings, we utilized Google's open building dataset [17]. Finally, OSM was again used to extract roads from the study area and measure their distance from forests. The amount of fuel biomass is estimated from Landsat's Normalized Difference Vegetation Index (NDVI). It is built from the red (R) and Near Infrared (NIR) channels, where:

$$\text{NDVI} = (NIR - R) / (NIR + R) \quad (9)$$

The dataset has the following information: Slope (percent), Aspect or Exposure (in degrees), elevation or Altitude (in meters), NDVI ranges between -1 to 1, Longitude and Latitude in degrees, Distance to

nearest road (in meters) and Distance to nearest dwelling area (in meters).

After data preprocessing, the data shape (rows x columns) = 80982 x 9.

The final result is obtained after calculating the fire risk according to the three models: fuzzy, Dargone and Erten.

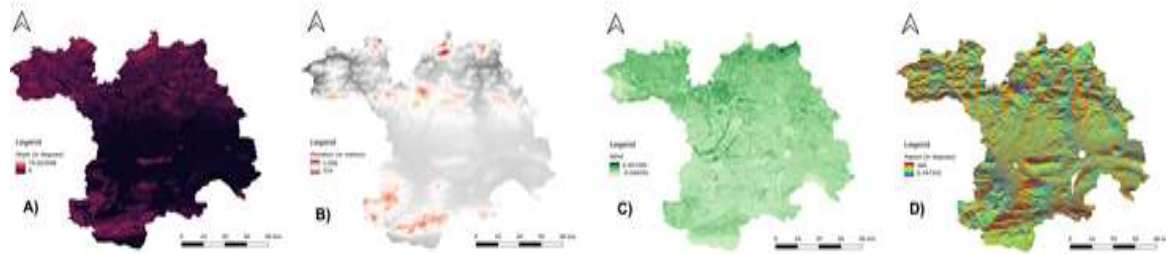


Figure 6. (A) Setif slope map (B) Setif Elevation map (C) Setif NDVI map (D) Setif aspect map

5.4 Results and discussion

As shown in Figure7, the results indicate a significant divergence between the Erten and Dargone methods with the fuzzy method. There is, however, a slight convergence between the Erten and Dargone methods.

According to the fuzzy method, 57.13% of the forest area was categorized as very high fire risk and 42.45% as high fire risk, with the remaining 0.42% of the study area at moderate fire risk. In contrast, the Dargone and Erten methods identified 32.13% and 25.77% of the area as very low fire risk, respectively, and 57.5% and 71.14% as low fire risk. The rest of the study area was found to be at moderate fire risk (10.37% for Dargone and 3.09% for Erten).

Reality shows that forest fires are increasing in the Setif region. It is a high-risk area where every year it is affected by this phenomenon. For example, in 2023, 42 fires that burned more than 113 hectares of forests and agricultural fields hit Setif. These fires notably broke out in the localities of Nawal Aït Mezada, Ain El Gueradj, Serdj El Ghoul, Beni Ouessine, Errasfa, and Beni Chebana (Agradou on the border with the Bejaia province).

According to [18], the number of forest fire outbreaks recorded in the Setif region between 1985 and 2004 varies between 11 (the minimum value recorded in 1995) and 75 (the maximum value recorded in 1987) with an average of 35 outbreaks per year. The burned area is considerably smaller compared to neighboring regions (Bejaia and Jijel). Over the twenty years from 1985 to 2004, this area reached 3314.69 hectares.

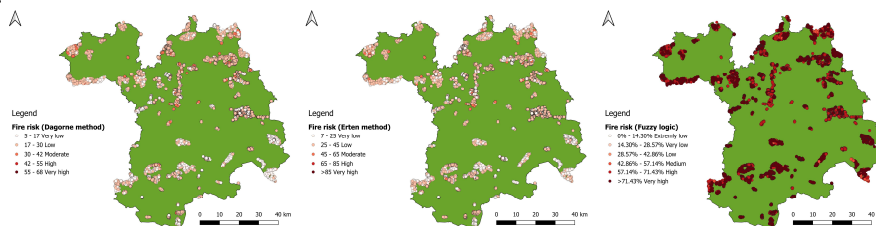


Figure 7. Fire risk according to: Fuzzy model, Dargone and Erten

6. Conclusion

All Mediterranean regions face devastating forest fires every year, resulting in human casualties, material damage, and ecosystem degradation. The Algerian case is symptomatic of this situation. The statistics available in Algeria, spanning a period of a century and a half, show an increase in the frequency of fires over the years. The growing deployment of firefighting resources at different stages of these disasters has helped stabilize the affected areas and significantly reduce the damage.

It is crucial to identify high-risk regions to effectively focus efforts on fire prediction and control before they spread uncontrollably.

Our contribution through this work was to propose a fuzzy system that includes various factors estimated to be conducive to the risk of forest fires: slope, altitude, exposure, vegetation, and human presence. We would have liked to test our model across the entire Algerian territory. Unfortunately, the absence of an Algerian database forced us to limit our study to the Setif region.

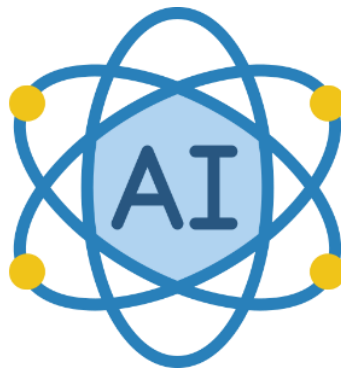
In the future, we plan to use deep learning to design a fire prediction model. It is also necessary to build a continuously updated dataset on the national forest cover to adapt to changes due to fire damage or reforestation.

References

- [1] The world bank, “New Study Suggests Ways to Boost Algeria’s Disaster Resilience”, ALGIERS, December 12, 2023. <https://www.worldbank.org/en/news/press-release/2023/12/12/new-study-suggests-ways-to-boost-Algeria-s-disaster-resilience>
- [2] O. Meddour-Sahar, R. Meddour, A. Derridj, “Analyse des feux de forêts en Algérie sur le temps long (1876-2007)”, Les Notes d’analyse du CIHEAM(Centre International de Hautes Etudes Agronomiques Méditerranéennes), vol. 39, September 2008.
- [3] A. Madoui, “Les incendies de forêt en Algérie. Historique, bilan et analyse. Forêt Méditerranéenne”, XXIII (1), pp. 23-30, hal-03558711, 2002
- [4] O. Meddour-Sahar, R. Meddour, V. Leone, R. Lovreglio, A. Derridj, “Analysis of forest fires causes and their motivations in northern Algeria: the Delphi method”, iForest, vol. 6, pp. 247-254, June 13th 2013. Ilham Loucif-Kabouya, “Enjeux des feux de forêt en Algérie”, Directorate of Wildlife and Flora Protection, General Directorate of Forests, Ministry of Agriculture and Rural Development, Algeria, 2021
- [5] J.K. Agee, “Fire Ecology of Pacific Northwest Forests”, Island Press, Washington, D.C. / Covelo, California, 1996
- [6] S. Aouda, Y. Mahdi, “Indice de risque d’incendie de forêts à l’aide du SIG dans la wilaya de Skikda, Algérie”, Revue Agrobiologia, ISSN (Print): 2170-1652, e-ISSN (Online): 2507-7627, vol. 13(2), pp. 3664-3675, 2023
- [7] H. Baazi, B. Labchaa, M. Naili, N. Drifi, “Cartographie du risque feu de forêt dans le Nord algérien: cas de la forêt d’Akkadou Est (W. de Bejaia)”, Algeria Geomatics Innovation Journal (AGIJJournal), ISSN: 2830-9359, Vol. 1, N° 1, 2022.
- [8] U. Nodelman, C. Allen, J. Perry, “Fuzzy Logic”, Stanford Encyclopedia of Philosophy, Bryant University, 2008.
- [9] O. MEDDOUR-SAHAR, R. MEDDOUR, A. Derridj, “Les facteurs favorables aux incendies de forêts en région méditerranéenne”, Faculté des Sciences Biologiques et des Sciences Agronomiques, Mouloud Mammeri university, Tizi Ouzou, Algeria, Revue campus n°17, pp. 4-19, 2010.
- [10] J. Olabarria, M. Palahi, A. Trasobares, and T. Pukkala, Timo, “A fire probability model for forest stands in Catalonia (north-east Spain)”, Annals of Forest Science, Vol. 64, pp. 169-176, February 2007. DOI: 10.1051/forest:2005109
- [11] J. Godjevac, “Idées nettes sur la logique floue”, Presses Polytechniques et Universitaires Romandes, Lausanne, Switzerland, 1999.
- [12] L. Gacogne, “Logique floue et applications”, In: Institut d’Informatique d’Entreprise, Evry, France, 2001.
- [13] K. Zerroug, “Elaboration d’un système d’information géographique (flore) dans la Wilaya de Sétif”, Master’s dissertation, Option: Biodiversity and ecosystem management, University Ferhat Abbas Setif1, Algeria, June 30th 2012
- [14] OpenStreetMap contributors,” Planet Dump” retrieved from <https://planet.openstreetmap.org>, 2024.
- [15] NASA Shuttle Radar Topography Mission (SRTM) (2013). <https://doi.org/10.5069/G9445JDF>.
- [16] W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y.S.E. Bouchareb, Y. Dauphin, D. Keyzers, M. Neumann, M. Cisse, J.A. Quinn. Continental-scale building detection from high resolution satellite imagery. ArXiv: 2107.12283, 2021.
- [17] Saifi, Merdas, “Bilan des incendies de forêts dans quelques wilayas de l’Est algérien; cas de Bejaia, Jijel, Sétif et Bordj Bou-Arredidj”, 10.13140/RG.2.2.25095.70567, 2007

SESSION 3

AI and Machine Learning Applications



Leveraging Transfer Learning and Data Augmentation for Enhanced Mental Health systems in Chronic Disease Management

Embarki Omar¹, Laifa Merieme², Mohdeb Djamila³

¹ Department of Computer Science, University of Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj, Algeria
(embarki24@gmail.com)

^{1,2} Laboratory of Informatics and Its Applications of M'sila (LIAM), M'sila, Algeria (meriem.laifa@univ-bba.dz)
<https://orcid.org/0000-0002-2010-9527>

³ Department of Computer Science, University of Mohamed Seddik Benyahia, Jijel, Algeria
Laboratory of Intelligent Systems and Cognitive Computing (ISCC)
(djamila.mohdeb@univ-jijel.dz)

Abstract

Patients with chronic diseases in developing countries like Algeria face increased mental health risks and limited access to care. AI presents new opportunities for managing and monitoring mental health, particularly through the use of sentiment analysis in digital communications. However, research on Transfer Learning (TL) models for Algerian dialects remains limited, compounded by a shortage of comprehensive datasets in the mental health domain. This paper aims to bridge this gap by focusing on Algerian Arabic, conducting experiments to annotate a dataset, and evaluating the performance of three TL models. Additionally, we implemented text augmentation techniques, demonstrating that pretrained language models tailored for Arabic, when combined with data augmentation, significantly enhance sentiment classification of user-generated content.

Keywords: Transfer learning, Sentiment analysis, Mental health, Chronic Diseases, Data augmentation

1. Introduction

We, Algerians, are facing a serious morbidity expansion; a phenomenon in which the population has longer lifespans but is accompanied by an increase in poor health [1]. The number of chronic diseases in Algeria has increased sharply in the recent few years due to changes in lifestyle, such as poorer diets, less physical activity, and rising obesity rates, as well as an aging population and better diagnostic and screening methods [2]. With the increase of chronic illness, comes an increase in mental disorders. The relationship between chronic diseases and mental health in Algeria presents a complex and often underestimated challenge. While specific data for Algeria is limited, research in developing countries suggests a significant prevalence of mental disorders among individuals with chronic physical illnesses [3]. This underscores the substantial burden mental health issues place on those with chronic conditions. The interplay between mental and physical health is bidirectional.

Despite recent improvements, Algeria's mental health care system faces major challenges. The treatment gap is widening as the prevalence of mental disorders rises faster than the growth of services. Access remains uneven, with services concentrated in the northern region, and the system

largely depends on a hospital-centered model, neglecting prevention and community integration [4]. Issues include insufficient resources, a lack of community-based care, poor service integration, and persistent stigma [5]. The country also struggles with limited mental health professionals, outdated approaches, scarce epidemiological data, legislative gaps, and implementation difficulties [5]. Artificial Intelligence (AI), particularly through Natural Language Processing (NLP) and Transfer learning (TL), has the potential to significantly improve diagnostic accuracy, predict and monitor public health surveillance across various conditions, personalize treatment plans, optimize resource allocation, and expand access to healthcare, particularly in underserved areas [6]. However, integrating such systems into the Algerian mental health system faces significant challenges, primarily due to two factors: the linguistic complexities of Algerian vernaculars and the scarcity of appropriate Algerian datasets. Although recent studies have leveraged advancements in transfer learning to enhance sentiment analysis of Algerian Arabic, including its dialects, the continued absence of suitable datasets remains a major obstacle to developing effective platforms for monitoring mental health among patients. Data augmentation techniques can help mitigate this issue by generating additional training data, which is essential for training models. This is particularly beneficial in the context of Algerian Arabic sentiment analysis, where collecting large, labeled datasets is challenging.

The purpose of this study is to address this gap by building the foundation stone towards integrating AI tools in the Algerian mental health systems for people with chronic illnesses. In particular, we focus on conducting a series of experiments involving the comparison of transfer learning models' performance and the impact of text augmentation techniques on them. The remainder of this paper is organized as follows: Section 2 describes the important role of AI and NLP in mental health systems in general. Our methodology is described in Section 3, followed by Section 4 in which the results of the experiments are presented and discussed. Section 5 concludes the paper.

2. AI and NLP mental health system

AI is transforming mental health research, diagnosis, and treatment by using ML and NLP techniques to analyze large datasets from electronic health records and social media. These methods help psychologists uncover patterns and markers of mental health conditions with greater speed and accuracy, including detecting subtle changes in language, facial expressions, and vocal tones [7]. Additionally, AI-driven systems provide continuous monitoring and support by integrating wearable technology and mobile health apps, alerting individuals to signs of distress, which is particularly beneficial for those with chronic mental health issues [8].

NLP can also be used to transcribe and analyze therapy sessions, giving therapists deeper insights into patients' speech patterns, emotional states, and underlying issues, which can lead to more personalized and effective treatment plans [9]. Moreover, NLP can be used to create predictive models that can identify individuals at risk of mental health crises. By analyzing large datasets of patient interactions, these models can detect subtle linguistic patterns and changes that precede a crisis, enabling timely intervention and potentially preventing severe mental health episodes [10]. As technology continues to evolve, the role of NLP in therapeutic settings is expected to become even more sophisticated, offering greater benefits to both patients and mental health professionals.

However, the integration of AI and NLP in mental health systems in Algeria remains largely unexplored in academic literature despite the Algerian government's push for healthcare advancements using new technologies. One major challenge is the lack of adequate datasets. Two main solutions to this problem are pre-trained TL models and data augmentation techniques. Research has demonstrated

that combining these approaches can significantly improve model performance, reduce the need for extensive labeled data, and enhance generalization across various domains, including sentiment analysis [11]. For example, Prades [12] conducted a comprehensive study on the application of large language models (LLMs) and data augmentation techniques for binary sentiment classification using a small dataset. The research highlighted that leveraging "nlpaug" for synonym replacement and GPT-2 for generating synthetic data notably enhanced the performance of LLMs through the processes of fine-tuning and soft prompting.

In a different study, Omran et al. [13] investigated the impact of data augmentation on sentiment analysis of translated texts across different languages, including English, Modern Standard Arabic (MSA), and the Bahraini dialect. By utilizing the "random swap" technique in conjunction with an LSTM model, their findings revealed significant improvements in accuracy and generalizability, with particularly notable effects observed in low-resource languages. Similarly, Badri et al. [14] explored the use of pre-trained language models (PLMs) and data augmentation for the detection of Arabic hate speech. Their research demonstrated that data augmentation techniques substantially improved classification metrics across various Arabic dialects. Although these techniques have shown considerable success in various domains, there is still a lack of exploration into their application in the context of mental health monitoring for chronic disease patients. This study aims to fill that gap by examining how these methods can be adapted for this specific and crucial area of research.

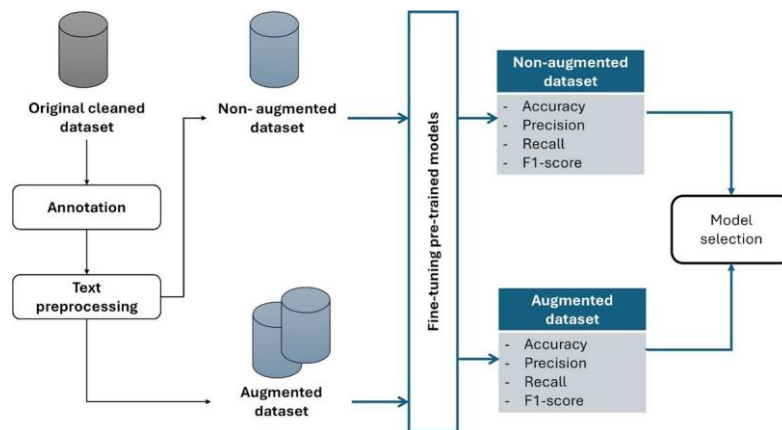


Figure 1. The proposed methodology

3. Methodology

This study integrated transfer learning and data augmentation to enhance sentiment analysis, focusing on dataset balance and model robustness as demonstrated in Figure 1. Our experimental design involved comparing models performance with and without data augmentation, and assessing the effects of freezing versus unfreezing pre-trained model layers to understand the trade-offs between adaptability and overfitting.

We used a dataset collected by Arbouche and Adjout [15] consisting of 3562 Facebook comments from Algerian chronic disease groups. After cleaning for duplicates, errors, and missing data, non-Arabic comments were translated to Arabic, reducing the dataset to 2981 comments. The cleaned dataset was annotated in two phases: two annotators initially labeled the comments as positive or negative, and a third annotator resolved any disagreements, finalizing the labels. After annotation, we observed an unbalanced label distribution of 10% negative and 90% positive instances. To enhance

model performance, we performed essential preprocessing steps to clean and standardize the text. This involved letter normalization, removing specific symbols, emojis, digits, and repeated letters, as well as eliminating unknown characters (non-Arabic), URLs, diacritics, punctuation, and user mentions.

We implemented three pre-trained Arabic language models: AraBERT [16], AraELECTRA [17], and DziriBERT [18]. Our study explored the efficacy of these models under different conditions, comparing their performance with frozen and unfrozen base layers.

A key aspect of our methodology was the application and comparison of several data augmentation techniques. We employed standard methods including random insertion, random deletion, word shuffling, and text masking. Additionally, we introduced a novel augmentation technique (See Figure 2) designed to enhance text data for training models. This technique involves breaking the text into 4-word groups (4-grams) and identifying the 25 most common sequences. These 4-grams are then randomly inserted back into the original text to create variations. Each original sequence undergoes this process five times, resulting in multiple augmented versions. This method is inspired by the random insertion technique.

Algorithm : Extracting 4-grams Sequences and Perform Augmentation

```
1: function EXTRACT4GRAMSEQUENCES(text_data)
2:   flattened_text_data ← [token for seq in text_data for token in seq]
3:   four_grams ← [(flattened_text_data[i : i + 4]) for i in ((flattened_text_data) - 3)]
4:   return four_grams
5: end function
6: function EXTRACTTOP254GRAMS(four_grams)
7:   four_gram_counts ← COUNTER(four_grams)
8:   top_25_four_grams ← four_gram_counts.MOST_COMMON(25)
9:   return [gram for gram, count in top_25_four_grams]
10: end function
11: function PERFORMAUGMENTATION(original_data, top_25_4_grams, num_insertions)
12:   augmented_data ← []
13:   for row in original_data do
14:     augmented_row ← row.COPY()
15:     for _ in (num_insertions) do
16:       insertion_index ← RANDOM.RANDINT(0, (augmented_row))
17:       selected_gram ← RANDOM.CHOICE(top_25_4_grams)
18:       AUGMENTED_ROW.INSERT(insertion_index, selected_gram)
19:     end for
20:     AUGMENTED_DATA.APPEND(augmented_row)
21:   end for
22:   return augmented_data
23: end function
```

Figure 2. The proposed augmentation algorithm

We conducted two experiments for each model: one with frozen base layers, training only the

classification head, and another with unfrozen layers for full fine-tuning. Both experiments used identical settings (5 epochs, batch size of 8, and gradient accumulation step of 4), managed by the Trainer class from Transformers. These experiments were performed first on the original data and then with augmented data to assess the impact of data augmentation. The experiments were run on a Kaggle environment with GPUs, using accuracy, precision, recall, and F1-score for evaluation.

3. Results and discussion

Each one of the following tables provides the evaluation results of the three transfer models on the two versions of the datasets (before and after text augmentation). In all results, the bold numbers represent the best scores and the underlined represents the second best. **Table 1** presents the evaluation metrics for transformers of each class using the original imbalance dataset, while **Table 2** reports the macro-average classification results. The results consistently showed that models with frozen base layers performed adequately on the positive class but were significantly deficient in predicting the negative class, leading to extremely low metrics for precision, recall, and F1-score for this class. This outcome underscores the limitations of frozen models in capturing the nuanced patterns necessary for balanced classification.

Table 1. Classification report for pre-trained models without data augmentation

		Negative			Positive		
	<i>TL Models</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
Base layer frozen	Base-AraBert	0.00	0.00	0.00	0.91	1.00	0.96
	AraElectra	0.00	0.00	0.00	0.91	1.00	0.96
	DziriBert	0.00	0.00	0.00	0.91	0.99	0.95
Base layer freed	Base-AraBert	0.48	0.24	0.32	0.93	0.98	0.95
	AraElectra	0.48	0.22	0.31	0.93	0.98	0.95
	DziriBert	0.59	0.41	0.48	0.95	0.97	0.96

Table 2. Macro-average report of pre-trained models performance without data augmentation

	<i>TL Models</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
Base layer frozen	Base-AraBert	0.91	0.46	0.50	0.48
	AraElectra	0.91	0.46	0.50	0.48
	DziriBert	0.91	0.46	0.50	0.48
Base layer freed	Base-AraBert	0.92	0.77	0.61	0.64
	AraElectra	0.91	0.70	0.60	0.63
	DziriBert	0.91	0.77	0.69	0.72

Unfreezing the base layers led to marked improvements in the models' performance on the negative class. These improvements suggest that allowing the base layers to adapt to the specific characteristics of the dataset enables the model to learn more effectively from both classes, thus achieving a more balanced classification. This finding is particularly important for small and imbalanced datasets, where the ability to adapt is crucial for achieving robust performance. However, the benefits of unfreezing the base layers come with significant drawbacks, most notably the increased risk of overfitting. This was evident in the observed discrepancies between training and validation losses, as well as in the Precision-Recall Curves (PRC) and Receiver Operating Characteristic (ROC) curves¹. The models, while showing high performance metrics initially, tended to overfit, particularly in the absence of data augmentation. This overfitting highlights the vulnerability of transfer learning models when their base layers are unfrozen, as they may become overly attuned to the training data, thereby reducing their generalizability to new data.

¹ Due to the limited pages, all figures can be found on github.com/embarki34/ISIA-24

Table 3. Macro-average report for transformers after data augmentation with the base layer frozen

Augmentation strategy	TL Models	Accuracy	Precision	Recall	F1-score
Random Insertion	Base-AraBert	0.76	0.75	0.75	0.75
	AraElectra	0.89	0.90	0.89	0.89
	DziriBert	0.71	0.71	0.71	0.71
Random Deletion	Base-AraBert	0.71	0.71	0.71	0.71
	AraElectra	0.70	0.70	0.70	0.70
	DziriBert	0.71	0.71	0.71	0.71
Word Shuffling	Base-AraBert	0.73	0.73	0.73	0.73
	AraElectra	0.78	0.78	0.77	0.77
	DziriBert	0.74	0.74	0.74	0.74
Word Masking	Base-AraBert	0.70	0.70	0.70	0.70
	AraElectra	0.76	0.76	0.76	0.76
	DziriBert	0.74	0.74	0.74	0.74
Our strategy	Base-AraBert	0.85	0.86	0.85	0.85
	AraElectra	0.83	0.84	0.82	0.82
	DziriBert	0.86	0.87	0.86	0.86

Table 4. Macro-average report for transformers after data augmentation with the base layer freed

Augmentation strategy	TL Models	Accuracy	Precision	Recall	F1-score
Random Insertion	AraBert	0.98	0.98	0.98	0.98
	AraElectra	0.99	0.99	0.99	0.99
	DziriBert	0.98	0.98	0.98	0.98
Random Deletion	AraBert	0.95	0.95	0.95	0.95
	AraElectra	0.94	0.94	0.94	0.94
	DziriBert	0.95	0.95	0.95	0.95
Word Shuffling	AraBert	0.99	0.99	0.99	0.99
	AraElectra	0.98	0.98	0.98	0.98
	DziriBert	0.98	0.98	0.98	0.98
Word Masking	AraBert	0.99	0.99	0.99	0.99
	AraElectra	0.97	0.97	0.97	0.97
	DziriBert	0.99	0.99	0.99	0.99
Our strategy	AraBert	1	1	1	1
	AraElectra	1	1	1	1
	DziriBert	1	1	1	1

The analysis across different implemented data augmentation techniques further illuminated the complex dynamics at play. **Table 3** presents the macro-average report for transformers results after data augmentation with the base layer frozen, while **Table 4** represents the macro-average report for transformers after data augmentation with the base layer freed. Based on the results, models with frozen base layers consistently underperformed, with macro average scores ranging from 0.64 to 0.89, depending on the method. This pattern reinforces the idea that freezing limits the model's ability to adapt and learn from new, diverse data, leading to suboptimal performance. However, our

augmentation strategy provided a significant improvement to all three models compared to traditional augmentation techniques, with the exception of the AraElectra model that performed better when applying random insertion technique. We believe that our method was efficient due to the fact that it added relatively meaningful sentences instead of random words.

On the other hand, unfreezing the base layers significantly improved performance, with scores consistently above 0.90 for the unfrozen models, demonstrating the clear advantage of adaptability in transfer learning models. Despite the improved performance with unfrozen layers, the recurring issue of overfitting indicates a critical trade-off: while unfreezing enhances the model's ability to fine-tune its parameters and better capture the dataset's characteristics, it also heightens the risk of overfitting, particularly in datasets that are small and imbalanced. This trade-off suggests that while unfreezing base layers is beneficial for performance, it must be applied with caution. Incorporating strategies to prevent overfitting, such as using regularization techniques or incorporating robust data augmentation methods, is essential to balance the improved adaptability with the need for generalizability.

5. Conclusion

The purpose of this study was to provide the first steps towards integrating AI into mental health systems. Our methodology allowed for a thorough investigation of advanced sentiment analysis techniques in the context of Algerian Arabic text, addressing the challenges of data scarcity and datasets imbalance in the mental healthcare field. A key aspect of our methodology was the application and comparison of several data augmentation techniques. The results highlight the importance of carefully managing the base layers in transfer learning models. Unfreezing the base layers offers substantial performance gains by enhancing the models' adaptability, but it also introduces the risk of overfitting, which must be mitigated to ensure the models' effectiveness on unseen data. This nuanced understanding of the trade-offs involved provides valuable guidance for optimizing transfer learning models in the context of small and imbalanced datasets. In future work, we aim to expand the dataset and diversify its sources to enhance the model robustness and generalizability. Additionally, we plan to conduct a comparative study with emerging augmentation techniques to establish best practices for low-resource languages and multilingual datasets, as well as exploring the integration of contextual information in the augmentation process to better preserve semantic meaning in Algerian Arabic dialects. The resulting improved sentiment analysis model would be integrated into a mental health monitoring system after conducting pilot studies in collaboration with mental health professionals to validate the system's effectiveness in real-world scenarios.

References

- [1] F. Flici and M. Chinoune, "Analysis of recent changes in chronic disease-free life expectancy in Algeria," *East Mediterr Health J.*, vol. 28, no. 12, pp. 872–878, Dec. 2022, doi: 10.26719/emhj.22.091.
- [2] M. Chinoune, F. Flici, and G. Saidi, "Estimating and Forecasting Chronic Disease-Free Life Expectancy in Algeria," 2024. doi: 10.13140/RG.2.2.18385.34400.
- [3] L. O. Daré *et al.*, "Co-morbidities of mental disorders and chronic physical diseases in developing and emerging countries: a meta-analysis," *BMC Public Health*, vol. 19, no. 1, p. 304, Mar. 2019, doi: 10.1186/s12889-019-6623-6.
- [4] M. E. M. Serid and B. Belaadi, "État des lieux des troubles mentaux et de leur prise en charge en Algérie," *East Mediterr Health J.*, vol. 28, no. 7, pp. 532–538, Jul. 2022, doi: 10.26719/emhj.22.054.
- [5] Z. Benmebarek, "Mental health law in Algeria: new amendments, old concerns," *BJPsych International*, vol. 19, no. 1, pp. 21–24, Feb. 2022, doi: 10.1192/bji.2021.14.
- [6] N. Sheliemina, "The Use of Artificial Intelligence in Medical Diagnostics: Opportunities, Prospects and

- Risks,” *HEM*, vol. 5, no. 2, pp. 104–124, Jul. 2024, doi: 10.21272/hem.2024.2-07.
- [7] R. Dave, K. Sargeant, M. Vanamala, and N. Seliya, “Review on Psychology Research Based on Artificial Intelligence Methodologies,” *Journal of Computer and Communications*, vol. 10, no. 5, Art. no. 5, May 2022, doi: 10.4236/jcc.2022.105008.
- [8] D. B. Olawade, O. Z. Wada, A. Odetayo, A. C. David-Olawade, F. Asaolu, and J. Eberhardt, “Enhancing mental health with Artificial Intelligence: Current trends and future prospects,” *Journal of Medicine, Surgery, and Public Health*, vol. 3, p. 100099, Aug. 2024, doi: 10.1016/j.glmedi.2024.100099.
- [9] N. Shek, E. Lee, G. Lee, H. Ko, and E. Lee, “Interdisciplinary Applications of AI: Dermatology and Psychology,” *Journal of Engineering Research and Reports*, vol. 26, no. 8, pp. 34–48, Jul. 2024, doi: 10.9734/jerr/2024/v26i81227.
- [10] M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, “Natural language processing for mental health interventions: a systematic review and research framework,” *Transl Psychiatry*, vol. 13, no. 1, pp. 1–17, Oct. 2023, doi: 10.1038/s41398-023-02592-2.
- [11] Z. Jiang, W. Zaheer, A. Wali, and S. A. M. Gilani, “Visual sentiment analysis using data-augmented deep transfer learning techniques,” *Multimed Tools Appl*, vol. 83, no. 6, pp. 17233–17249, Jul. 2023, doi: 10.1007/s11042-023-16262-4.
- [12] G. S. Prades, “Modelling Sentiment Analysis: LLMs and augmentation techniques,” 2023.
- [13] T. Omran, B. Sharef, C. Grosan, and Y. Li, “The Impact of Data Augmentation on Sentiment Analysis of Translated Textual Data,” in *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, Mar. 2023, pp. 1–4. doi: 10.1109/ITIKD56332.2023.10099851.
- [14] N. Badri, F. Kboubi, and A. Habacha Chaibi, “Abusive and Hate speech Classification in Arabic Text Using Pre-trained Language Models and Data Augmentation,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Aug. 2024, doi: 10.1145/3679049.
- [15] A. M. Issiakhem and A. Lotfi, “Analyse des sentiments et modélisation thématique concernant les maladies chroniques,” Thesis, 2023. Accessed: Aug. 14, 2024. [Online]. Available: <https://dspace.univ-bba.dz:443/xmlui/handle/123456789/3875>
- [16] W. Antoun, F. Baly, and H. Hajj, “AraBERT: Transformer-based Model for Arabic Language Understanding,” Mar. 07, 2021, *arXiv*: arXiv:2003.00104. doi: 10.48550/arXiv.2003.00104.
- [17] W. Antoun, F. Baly, and H. Hajj, “AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding,” Mar. 07, 2021, *arXiv*: arXiv:2012.15516. doi: 10.48550/arXiv.2012.15516.
- [18] A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, “DziriBERT: a Pre-trained Language Model for the Algerian Dialect,” Dec. 12, 2022, *arXiv*: arXiv:2109.12346. doi: 10.48550/arXiv.2109.12346.

Application of Deep Learning for the Automated Detection and Classification of Diabetic Retinopathy from High-Resolution Medical Images

Ghizlane KHABABA^{1✉}, Abdallah KHABABA², Saad HAROUS³, Mohamed Abdessamad KRACHE⁴, Chouaib MAIZA⁵

¹ PhD Student, Deptment of computer science ,Faculty of sciences, University Setif 1
Ghizlane.khababa@univ-setif.dz

² Professor, Deptment of computer science ,Faculty of sciences, University Setif 1
akhababa@univ-setif.dz

³Professor, Department of computer science College of Computing and Informatics
University of Sharjah
Harous@sharjah.ac.ae

⁴ Master student, Deptment of computer science ,Faculty of sciences, University Setif1
abdessamedkrr18@gmail.com

⁵ Master student, Deptment of computer science ,Faculty of sciences, University Setif1
chouaibmaiza10@gmail.com

Abstract

Diabetic Retinopathy (DR), a leading cause of preventable blindness, necessitates timely detection and intervention to preserve vision. Traditional DR screening methods, reliant on ophthalmologists, are labor-intensive and prone to delays, underscoring the demand for efficient automated solutions. In this study, we leveraged the APTOS dataset, containing 3662 labeled high-resolution fundus images, to train and evaluate deep learning models for classifying DR into five severity stages (0 to 4). Three state-of-the-art architectures—DenseNet121, MobileNet, and Xception—were implemented to extract features and classify images. Among these, DenseNet121 demonstrated superior performance, achieving an accuracy of 95% and a Cohen kappa score of 0.968, indicating high consistency with expert annotations. Furthermore, a software application was developed to integrate these models, providing a user-friendly interface for rapid and accurate DR diagnosis. This application equips healthcare professionals with a powerful tool for early detection, potentially reducing the burden on specialists and improving patient outcomes. The proposed system bridges the gap between cutting-edge AI techniques and clinical usability, offering a scalable solution for DR management.

Keywords: Diabetic Retinopathy(DR), Deep Learning, Fundus Images classification, Automated Screening

1. Introduction

Healthcare is undergoing a transformative shift driven by technological advancements and a focus on early disease detection, particularly in managing chronic conditions like diabetes. Diabetes, which has reached epidemic levels globally, often leads to diabetic retinopathy (DR), a progressive eye disease that can cause blindness if undiagnosed and untreated [1]. Given the growing prevalence of diabetes, there is an urgent need for efficient DR screening methods to detect the disease early and prevent its severe consequences [2].

—Traditionally, DR screening is performed manually by ophthalmologists, a time-consuming and resource-

intensive process. However, advances in artificial intelligence (AI), particularly in machine learning and deep learning, offer promising solutions for enhancing diagnostic accuracy and efficiency. Convolutional Neural Networks (CNNs) [3], designed for image analysis, have shown great potential in DR detection

by automatically identifying features in retinal images, such as lesions or hemorrhages.

This study aims to enhance DR detection using CNN models like DenseNet121, MobileNet, and Xception, trained on a large dataset of retinal fundus images. These models have been integrated into a software application to assist healthcare professionals with quick and accurate diagnoses.

Despite progress, challenges remain, including the scalability of models with increasing data, maintaining a balance between accuracy and computational efficiency, and adapting to diverse datasets. Addressing these challenges is critical for practical applications in real-world healthcare.

This research contributes to improving DR detection by reviewing recent advancements in CNN-based approaches and highlighting the benefits of integrating AI-driven tools into clinical practice, ultimately aiming for better patient care and outcomes.

The remainder of this paper is organized as follows: Section 2 highlights the latest advances in DL and ML related to DR detection. Section 3 explains the steps of the study that would be implemented in section 4, this later highlights and discusses the obtained results. The paper is concluded in 5.

2. Related works

The use of deep learning for automated detection and classification of diabetic retinopathy (DR) from high-resolution medical images has garnered significant attention. Numerous studies have demonstrated the potential of deep learning algorithms to enhance diagnostic accuracy and efficiency in detecting DR, a common diabetes complication that can lead to vision loss if untreated [4]–[6]. These models have shown high sensitivity and specificity in analyzing retinal fundus photographs, achieving near-ophthalmologist-level accuracy in categorizing DR [5], [7], [8].

Researchers have integrated deep learning with techniques like image processing, neural networks, and feature extraction to improve detection and classification [9]–[11]. Additionally, segmentation methods for retinal blood vessels have been explored to optimize the performance of automated systems [9]. Recent advancements also include computer-aided systems using neural networks and ensemble modules for precise DR diagnosis [12].

Moreover, techniques such as transfer learning and attention mechanisms have been employed to improve model accuracy and interpretability [13]–[15]. These advancements aim to enhance DR screening and diagnosis in real-world settings, providing scalable, AI-driven solutions for early detection in primary care and large-scale screening programs. Overall, deep learning offers transformative potential in improving DR diagnosis, ultimately benefiting patients and healthcare systems.

3. Methodology

Experience and results are a key factor in advancing DR detection using deep learning. As researchers apply these models in clinical settings, they face challenges, gain insights, and achieve outcomes that drive progress in medical science.

3.1. Preprocessing

Using OpenCV, we preprocessed the diabetic retinopathy dataset by applying techniques like resizing, normalization, and contrast adjustment to enhance image quality and aid feature extraction. OpenCV enabled key tasks such as image cropping, noise reduction, and color correction, optimizing the dataset for model training. The goal was to standardize image characteristics, minimize noise, and highlight relevant features, ultimately improving the performance of machine learning models in DR diagnosis.

These steps are followed during the preprocessing (Fig 1 shows the images before and after the processing):

- 1) **Read the image:** Load the input image using the cv2.imread() function.
- 2) **Convert to blue tint:** Modify the color channels of the image to emphasize blue. Set the blue channel to maximum intensity while keeping the other channels unchanged as the blue channel could improve the model's ability to detect patterns or anomalies.
- 3) **Normalize the image:** Scale the pixel values of the image to a range between 0 and 1 by dividing by 255. This step ensures that all pixel values are within the same range, making it easier for the model to learn.
- 4) **Expand dimensions:** If necessary, expand the dimensions of the image to match the input shape expected by the model. This usually involves adding a new dimension for the batch size and possibly the channel dimension if working with a CNN.
- 5) **Output the preprocessed image:** The final preprocessed image is ready to be fed into your ML model for further analysis or training.

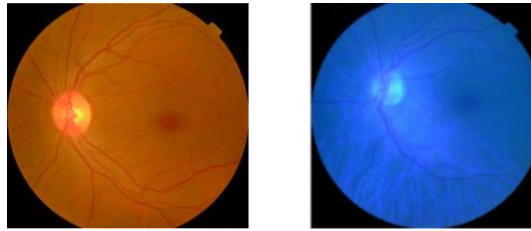


Figure. 1: Original and Preprocessed image

3.3. Methodology

In our study, we meticulously designed and implemented three distinct convolutional neural network (CNN) architectures tailored for the classification of retinal images extracted from the APTOS Blindness Detection dataset. These models were meticulously trained and evaluated to discern their efficacy in accurately identifying diabetic retinopathy severity levels.

As we know that DR detection problem is a primary step to detect blindness. To overcome this problem early detection is the first concern. So for early detection, we are using the deep learning architectures called “DenseNet 121 Architecture”, “MobileNet Architecture” and “Xception Architecture” (Figure 2).

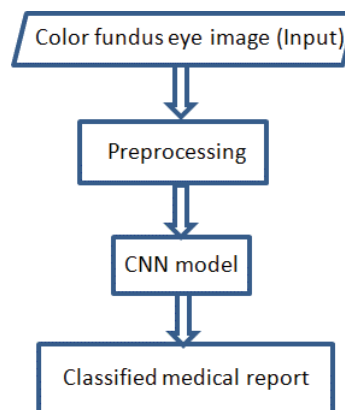


Figure. 2: Deep learning framework

3.3.1. DenseNet 121 Architecture

DenseNet models are designed to be more parameter-efficient compared to traditional convolutional neural networks (CNNs) by eliminating redundant feature maps. Instead of simply adding feature maps, DenseNet concatenates the outputs of each layer with the input feature maps, fostering dense connectivity within the network.

The dense connectivity in a DenseNet is mathematically represented by:

$$T_N = T_N + [Q_0, Q_1, \dots, Q_{N-1}]$$

where:

- T_N represents the output of the N^{th} layer.
- $[Q_0, Q_1, \dots, Q_{N-1}]$ denotes the concatenation of feature maps generated by the preceding layers.

DenseNets are organized into DenseBlocks, where the feature map dimensions remain constant within a block while the number of filters varies. This dense connectivity leads to an increase in the channel dimension with each layer is defined as follows:

$$R_N = R_0 + R \times (N - 1)$$

where:

- R represents the Rate.
- R_N represents the output of the N^{th} layer.

DenseNets increase the depth of DCNNs by reusing features, allowing for more efficient learning. In DenseNet121, fewer feature maps are required compared to other architectures. Unlike ResNet, which adds the outputs of layers, DenseNet concatenates them, enhancing feature propagation.

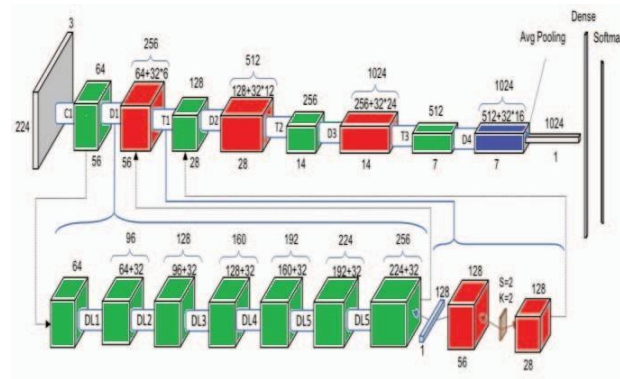


Figure. 3: DenseNet 121 Architecture

Figure 3 shows the block diagram of DenseNet 121 architecture: each volume measurement indicates the depth and width of the 2D features, while the top numbers denote featuredimensions. For example, a growth rate of 32 is used. The volume of each DenseBlock increases by multiplying the growth rate by the number of dense layers within the block. Each new layer adds to the previous feature set, resulting in an increase from 64 to 256 features after 6 layers. Furthermore, transition block performed as 1 X 1 convolution with 128 filters . 2 X 2 pooling with

a stride of 2, resulting in separating the size of the volume and the number of features on half.

The Cohan Kappa Score obtained for the proposed work is 96.8% and the Accuracy 95% (See Fig 4) for DR grading. The experiments and evaluation results show that the proposed method works well for accurate DR grading and early disease detection.

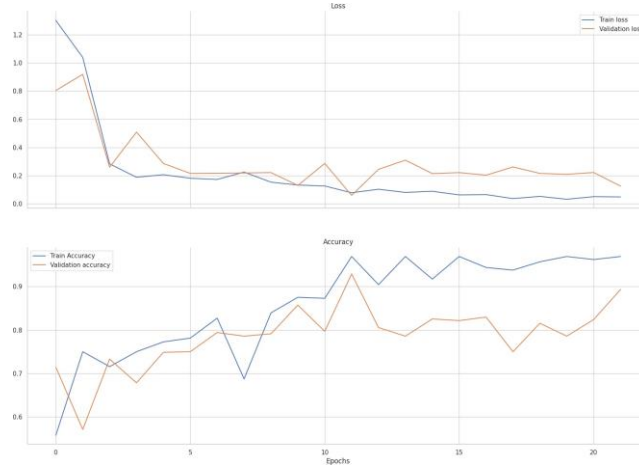


Figure. 4: Loss and Accuracy during training and validation

5. Experiments and evaluation

5.1 Dataset

The image data as shown in Figure 5 for this research was sourced from the Kaggle dataset provided by "Aravind Eye Hospital", as part of the 4th Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 Blindness Detection competition. This publicly available dataset, the largest of its kind, includes high-resolution retina images captured under various conditions using a fundus camera.

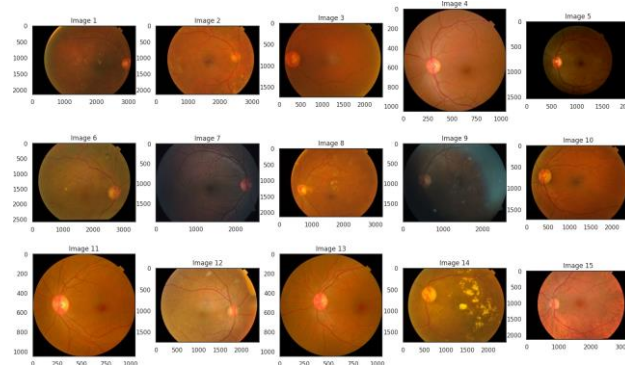


Figure. 5: Example of images in the APTOS 2019 database

The dataset consists of thousands of retinal images, each labeled with a severity grade from 0 to 4, representing the presence and progression of diabetic retinopathy (DR). The grades are as follows:

- Grade 0: No DR (No diabetic retinopathy): 1805
- Grade 1: Mild DR (Mild non-proliferative diabeticretinopathy): 370

- Grade 2: Moderate DR (Moderate non-proliferative diabetic retinopathy): 999
- Grade 3: Severe DR (Severe non-proliferative or early proliferative diabetic retinopathy) : 295
- Grade 4: Proliferative DR (Proliferative diabetic retinopathy): 193

These labels provide a comprehensive classification system for diagnosing and grading the severity of DR based on retinal images.

5.1. Experimental environment

- Computer: HP EliteBook 840 G5 Notebook PC
- RAM: 16 GB
- Processor: Intel(R) Core(TM) i5-8350U CPU @1.70GHz
- System Type: 64-bit operating system, x64 processor
- GPU: Intel(R) UHD Graphics 620

• Model Comparison to existing models

We evaluated our model's performance against existing classification methods, with the confusion matrix (shown in Figure 6) illustrating its accuracy in identifying diabetic retinopathy. The confusion matrix compares actual and predicted labels for five classes. Diagonal elements represent correct classifications, while non-diagonal elements indicate misclassifications.

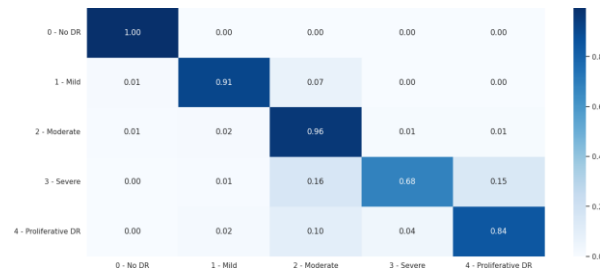


Figure. 6: Confusion Matrix of DenseNet121

5.2. MobileNet Architecture

MobileNet is a model constructed mainly from depthwise separable convolutions originally announced in and afterward used in Inception models to decrease the volume of computation constraints in the first few layers. Flattened networks construct a network out of fully factorized convolutions and present the potential of exceptionally factorized networks.

Factorized networks, like Xception, use depthwise separable convolutions to improve performance over Inception V3. SqueezeNet, a smaller network, uses a bottleneck design to reduce size, while other networks, like arranged transform networks and deep fried convnets, focus on reducing computations.

MobileNet is also built on depthwise separable convolutions, which split a regular convolution into two steps: depthwise convolution applies a filter to each input channel, and pointwise convolution (1×1) combines the outputs. This factorization significantly reduces both computation and model size.

MobileNet's key innovation lies in the use of depthwise separable convolutions, as it significantly reduces computational complexity while preserving representational capacity. Mathematically, a depthwise separable convolution can be represented as follows:

1)Depthwise Convolution :

$$G = (F \star K_d) \star K_p$$

where F is the input feature map, K_d is the depthwise convolution kernel, and G is the output feature map. The depthwise convolution operation $\star$ is followed by a pointwise convolution operation $\star$. K_p is the pointwise convolution kernel.

2)Pointwise Convolution :

In this step, the output of the depthwise convolution H_d is convolved with a pointwise convolution kernel K_p to produce the final output feature map G . The pointwise convolution is applied to mix information across channels. The pointwise convolution operation can be formulated as:

$$G = H_d \star K_p$$

* denotes the pointwise convolution operation.

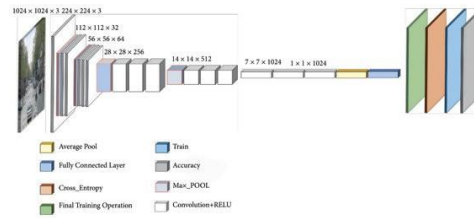


Figure. 7: MobileNet Architecture

The Cohen Kappa Score achieved by this model stands at 94.9%, while the accuracy reached 93%. (See fig 8)

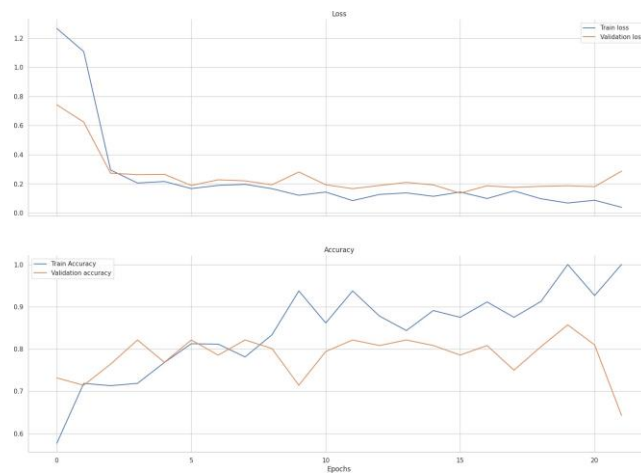


Fig. 8: Loss and Accuracy during training and validation

• Model Evaluation

These experiments have been conducted by retinal images through modal features that are obtained by our proposed dense deep features of CNN. Figure 9 shows identification levels of diabetic retinopathy through a confusion matrix.

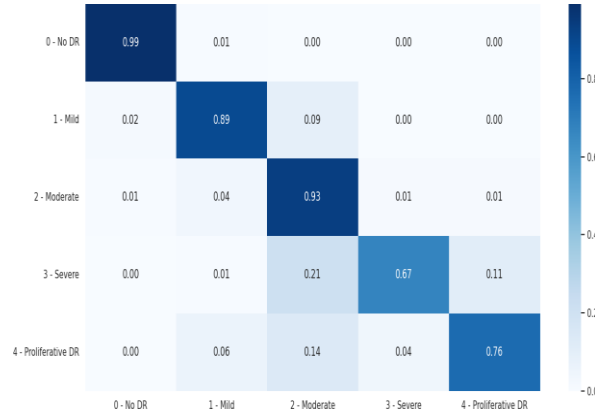


Figure. 9: Confusion Matrix of MobileNet

5.3. Xception Architecture

The Xception architecture, introduced by Francois Chollet, is an extension of the Inception architecture [16]. This architecture is a linear stack of depthwise separable convolution layers with residual connections. The depthwise separable convolution aims to reduce computational cost and memory requirements. Xception has 36 convolutional layers structured into 14 modules, all include linear residual connections, except for the first and last modules.

The separable convolution in Xception decouples the learning of channel-wise and spatial features. The residual connection, introduced by He et al. [18], addresses vanishing gradients and representational bottlenecks by creating a shortcut in the network. This shortcut sums the output of an earlier layer with a later one, instead of concatenating them. The Xception module employs depthwise separable convolutions in sequence.

Let F represent the input feature map, $K_1, K_2, \dots, K_n$ denote the depthwise separable convolution kernels, and G represent the output feature map. The Xception module is formulated as:

$$G = F \star K_1 \star K_2 \star \dots \star K_n$$

where $\star$ represents the depthwise separable convolution operation.

The Xception architecture is an extension of the Inception architecture. It replaces the standard convolutional layers with depthwise separable convolutions, which significantly reduces the number of parameters while maintaining performance. Next, we present a brief description of the Xception architecture:

Entry Flow:

- The input image is passed through a series of convolutional layers followed by batch normalization and ReLU activation functions.
- These initial layers serve to extract low-level features from the input image.

Middle Flow:

- The middle flow consists of a repeated sequence of depthwise separable convolutions.

Each depthwise separable convolution block is composed of three depthwise separable convolutional layers followed by batch normalization and ReLU activation functions.

- The middle flow helps to capture more complex features from the input.

Exit Flow:

- The exit flow is similar to the middle flow but includes additional operations to reduce the spatial dimensions of the feature maps.
- It consists of a series of depthwise separable convolutional layers followed by global average pooling to reduce the spatial dimensions to 1×1 .
- The output of the global average pooling layer is then passed through a fully connected layer with softmax activation to produce the final predictions.
- The Xception architecture has shown to achieve state-of-the-art performance on various computer vision tasks while being more computationally efficient compared to traditional architectures such as : VGG or ResNet.

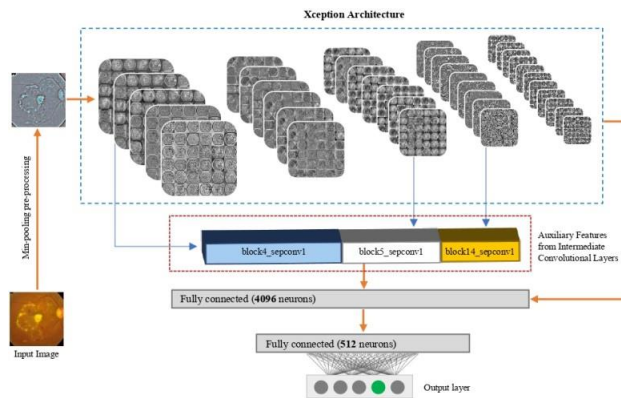


Figure. 10: Xception Architecture

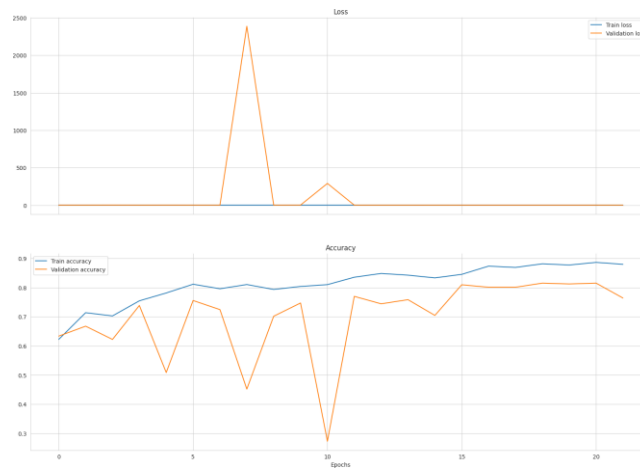
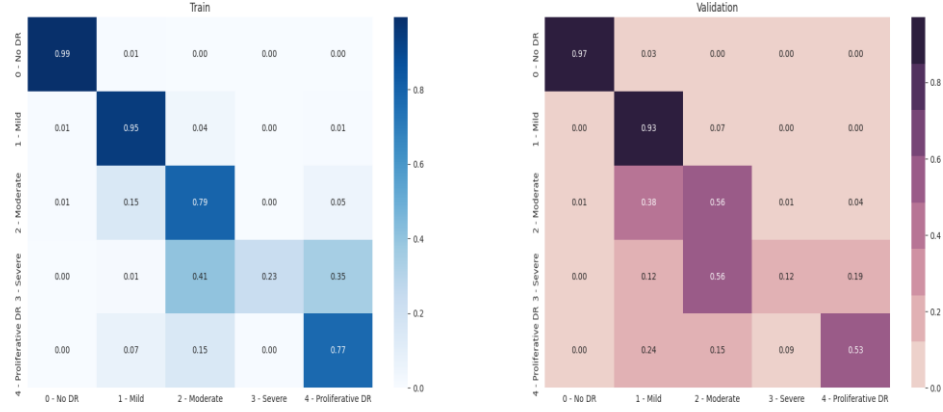


Figure. 11: Loss and Accuracy during training and validation

Model Evaluation

This examination has been conducted using retinal images, employing modal features

obtained by our pro-posed dense deep features of CNN. Fig 12 illustrates the identification levels of diabetic retinopathy through a confusion matrix specifically for the Xception model. This visualization showcases the model's performance in classifying diabetic retinopathy cases, providing valuable insights into its effectiveness in medical image analysis. Comparing these results to those obtained using other classification methods can further validate the robustness and accuracy of our proposed approach.



5.4. Classification Performance Measures

The performance measures are used to evaluate the performance of a binary classification model and calculate other performance metrics, such as: precision, recall, f1-score and accuracy.

- **True Positive (TP):** Number of normal data correctly classified as normal.
- **False Positive (FP):** Number of abnormal data incorrectly classified as normal.
- **True Negative (TN):** Number of abnormal data correctly classified as abnormal.
- **False Negative (FN):** Number of normal data incorrectly classified as abnormal.

5.5. Evaluation Metrics

Accuracy:

In order to know if the model is accurate, we must compute the precision which is an important evaluation measure and the most intuitive performance measure of observations correctly predicted in relation to the total number of observations.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Precision:

Precision is the ratio between the number of positive observations correctly predicted and the total number of positive observations predicted.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall:

Recall, called also sensitivity, is the relationship between positive observations correctly predicted and all observations of the real class. In other words, the recall measures the

proportion of positive real classes that are correctly identified

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- F1-score:

The F1 score can be interpreted as a weighted average of precision and recall, or an F1 score reaches its best value at 1 and its worst value at 0. If the cost of false positives and false negatives is very different, it is preferable to consider both precision and recall. F1 Unique metric score that combines recall and precision using harmonic mean.

$$\text{F1-score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Cohen Kappa score:

Cohen's Kappa score is a statistical measure that assesses the inter-rater agreement for qualitative items. It is commonly used in classification tasks to evaluate the agreement between two raters or models when assigning categorical labels to items.

The formula for Cohen's Kappa score is:

$$\text{kappa} = \frac{P_o - P_e}{1 - P_e}$$

After completing the experiments, we obtained results showcasing the validity of our proposed method. We tested three architectures on the same dataset to compare their accuracy.

Table1. Results and analysis

Architecture	Dataset	Kappa	Precision	Recall	F1-score	Accuracy
DenseNet121	APTOS 2019	96,8%	95%	95%	95%	95%
MobileNet	APTOS 2019	94,9%	93%	93%	93%	93%
Xception	APTOS 2019	Not used	80%	75%	75%	75%

5.6. DR GUI

We introduce the design and implementation of a Graphical User Interface (GUI) for diabetic retinopathy detection, leveraging the power of deep learning CNNs. This GUI is intended to provide healthcare professionals with a user-friendly tool to upload retinal images, process them through a trained CNN model, and receive diagnostic results quickly and accurately. The integration of CNNs within the GUI framework aims to enhance diagnostic precision, reduce the workload on medical practitioners, and facilitate early intervention. **Graphic User Interface** We present the interface of the application developed in this work.

The first interface This interface allows users to access the rest of the application.

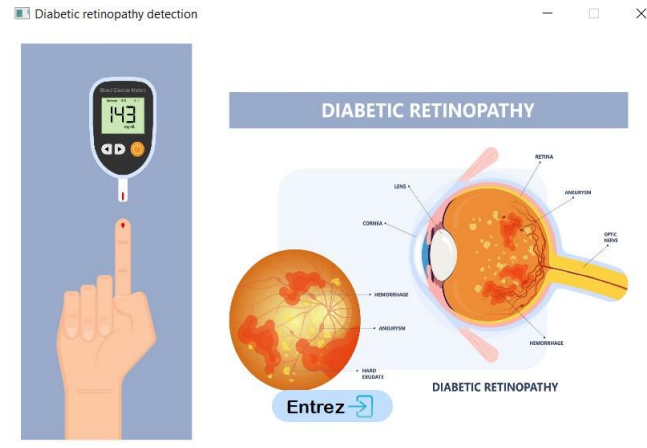


Fig. 13: The first interface of the application
The **Enter** button allows you to move to the next window.

The second interface

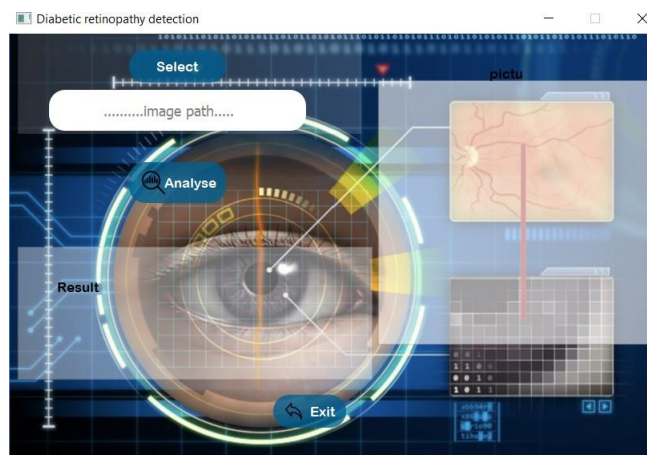


Fig. 14: Main interface

This interface is considered the main interface of the application. This tool allows you to upload and analyze images.

- **Select:** to upload an image.
- **Analyse:** for analyzing the image.
- **Exit:** to exit the application.

After clicking on the analysis button we have 5 possible cases:

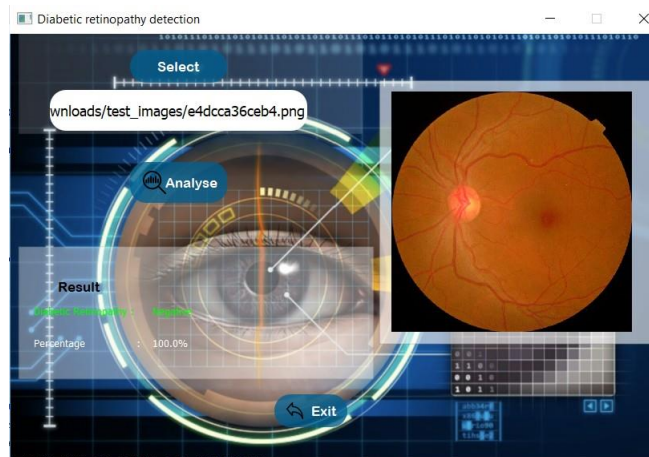


Fig. 15: Example of a No DR case.

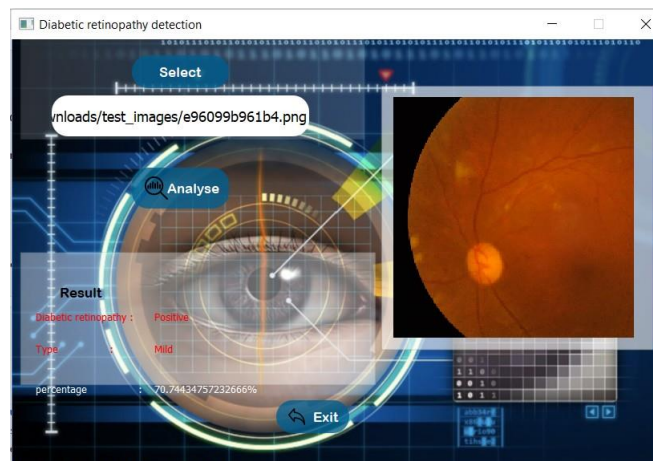


Fig. 16: Example of a mild DR case

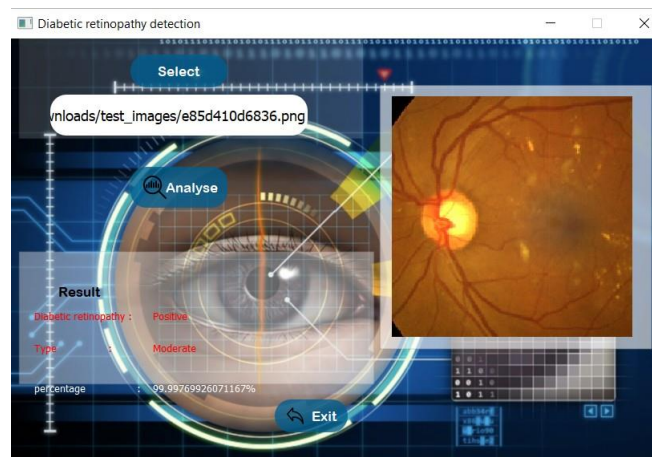


Fig. 17: Example of a moderate DR case

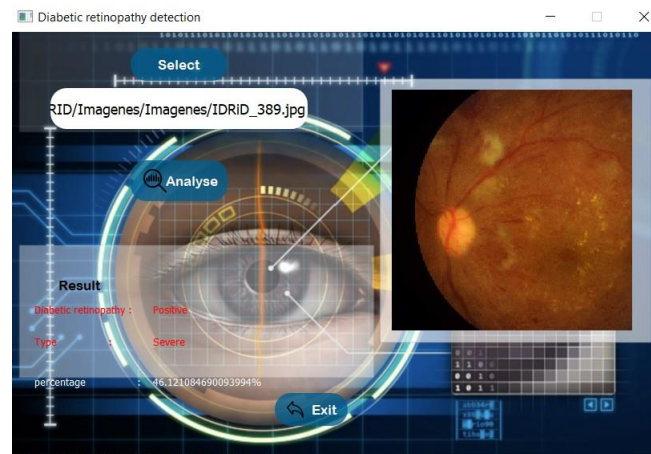


Fig. 18: Example of a severe DR case

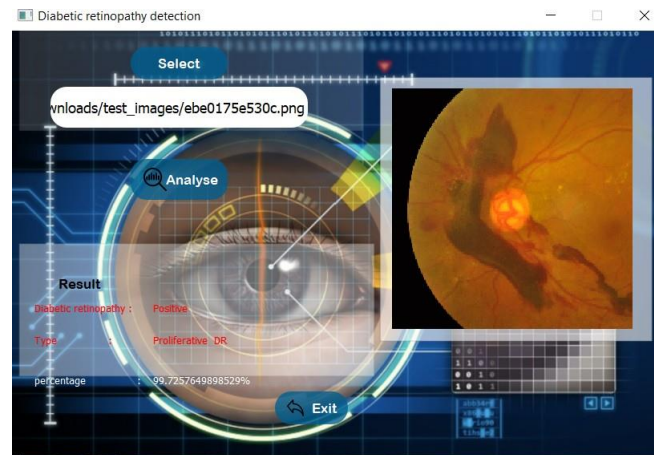


Fig. 19: Example of a Proliferative DR case

6. Conclusion

This study applied deep learning models, specifically Convolutional Neural Networks (CNNs), to improve the detection and classification of diabetic retinopathy (DR) using retinal fundus images. Leveraging the APTOS dataset of 3662 images, we trained DenseNet121, MobileNet, and Xception models to classify DR into five stages. DenseNet121 achieved the highest accuracy of 95% and a Cohen's kappa score of 0.968.

We also developed a user-friendly software application that integrates these models, allowing healthcare professionals to make rapid, accurate DR assessments. While promising, challenges related to model scalability and adaptability must be addressed for broader clinical adoption. Nonetheless, this work highlights the potential of automated DR detection to improve early diagnosis and patient outcomes.

Acknowledgements

The authors would like to thank Setif university and Sharjah University for providing support for this research.

References

- [1] Norah Asiri, Muhammad Hussain, Fadwa Al Adel, and Nazih Alzaidei. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artificial intelligence in medicine*, 99:101701, 2019.
- [2] Imran Qureshi, Jun Ma, and Qaisar Abbas. Recent development on detection methods for the diagnosis of diabetic retinopathy. *Symmetry*, 11(6):749, 2019.
- [3] Zhentao Gao, Jie Li, Jixiang Guo, Yuanyuan Chen, Zhang Yi, and Jie Zhong. Diagnosis of diabetic retinopathy using deep neural networks. *IEEE Access*, 7:3360–3370, 2018.
- [4] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.
- [5] Matthew S Hunt, Yuka Kihara, and Aaron Y Lee. Novel low-shot deep learning approach for retinal image classification with few examples. *JAMA ophthalmology*, 138(10):1077–1078, 2020.
- [6] Thippa Reddy Gadekallu, Neelu Khare, Sweta Bhattacharya, Saurabh Singh, Praveen Kumar Reddy Maddikunta, In-Ho Ra, and Mamoun Alazab. Early detection of diabetic retinopathy using pca-firefly based deep learning model. *Electronics*, 9(2):274, 2020.
- [7] Guanghua Zhang, Bin Sun, Zhixian Chen, Yuxi Gao, Zhaoxia Zhang, Keran Li, and Weihua Yang. Diabetic retinopathy grading by deep graph correlation network on retinal images without manual annotations. *Frontiers in Medicine*, 9:872214, 2022.
- [8] Mohammad B Khan, Mohiuddin Ahmad, Shamshul B Yaakob, Rahat Shahrrior, Mohd A Rashid, and Hiroki Higa. Automated diagnosis of diabetic retinopathy using deep learning: On the search of segmented retinal blood vessel images for better performance. *Bioengineering*, 10(4):413, 2023.
- [9] Muhammad Kashif Yaqoob, Syed Farooq Ali, Muhammad Bilal, Muhammad Shehzad Hanif, and Ubaid M Al-Saggaf. Resnet based deep features and random forest classifier for diabetic retinopathy detection. *Sensors*, 21(11):3883, 2021.
- [10] Yong Fu, Yuekun Wei, Siying Chen, Caihong Chen, Rong Zhou, Hongjun Li, Mochan Qiu, Jin Xie, and Daizheng Huang. Uc-stack: a deep learning computer automatic detection system for diabetic retinopathy classification. *Physics in Medicine & Biology*, 69(4):045021, 2024.
- [11] Yue Miao, Siyuan Tang, et al. Classification of diabetic retinopathy based on multiscale hybrid attention mechanism and residual algorithm. *Wireless Communications and Mobile Computing*, 2022, 2022.
- [12] G Kalyani, B Janakiramaiah, A Karuna, and LV Narasimha Prasad. Diabetic retinopathy detection and classification using capsule networks. *Complex & Intelligent Systems*, 9(3):2651–2664, 2023.
- [13] Se-In Jang, Michael JA Girard, and Alexandre H Thiery. Explainable and interpretable diabetic retinopathy classification based on neural-symbolic learning. *arXiv preprint arXiv:2204.00624*, 2022.
- [14] Sarah Meredith, Mark van Grinsven, Jonne Engelberts, Dominic Clarke, Vicki Prior, Jo Vodrey, Alison Hammond, Raja Muhammed, and Philip Kirby. Performance of an artificial intelligence automated system for diabetic eye screening in a large english population. *Diabetic Medicine*, 40(6):e15055, 2023.
- [15] Frank D Verbraak, Michael D Abramoff, Gonny CF Bausch, Caroline Klaver, Giel Nijpels, Reinier O Schlingemann, and Amber A van der Heijden. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. *Diabetes care*, 42(4):651–656, 2019.
- [16] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

Simplifying the DBSCAN Algorithm with a Single-Parameter Approach.

Makhlouf Benazi¹✉

¹ *Laboratory of Informatics and its Applications of M'sila (LIAM), University of M'sila, Department of computer science, University Pole, Road Bourdj Bou Arreiridj, M'sila 28000, Algeria
makhlouf.benazi@univ-msila.dz*

Abstract

This paper presents an enhanced version of the DBSCAN algorithm, termed myDBSCAN, which simplifies the clustering process by relying on a single parameter, `eps`, instead of the traditional two parameters used in the original DBSCAN. The study evaluates myDBSCAN's performance across various synthetic datasets with different shapes and levels of noise. Empirical results demonstrate myDBSCAN's performance comparably to the original DBSCAN algorithm, successfully identifying clusters with similar accuracy. The simplicity of using only one parameter makes myDBSCAN more accessible and easier to implement. The results highlighted myDBSCAN's effectiveness in clustering tasks, offering a practical alternative to traditional DBSCAN, particularly in scenarios where ease of parameter tuning is crucial. Future research will explore further optimizations and applications of myDBSCAN to a broader range of datasets.

Keywords: clustering, density-based algorithms, DBSCAN, Parameter optimization, eps parameter.

1. Introduction

In recent years, unsupervised learning, specifically data clustering, has emerged as a critical technique of machine learning. This method allows discovering meaningful patterns within datasets by grouping unlabeled data into subgroups called clusters. The primary goal is to ensure objects within the same cluster are as similar as possible, while objects in different clusters are as distinct as possible. [1].

Clustering has been applied across numerous domains, including astronomy [2], biology [3], computer science [4], insurance [5], and medicine [6]. Various clustering algorithms have been proposed in the literature, typically classified into four main categories: Partitioning Methods, such as k-means [7]; Hierarchical Methods, such as SLINK [8]; Model-Based Methods, such as Gaussian mixture models [9]; and Density-Based Methods, such as DBSCAN [10] and OPTICS [11].

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [10] is a well-established density-based clustering algorithm known for its ability to identify arbitrarily shaped clusters and resist noise interference. However, its effectiveness depends on two parameters, `epsilon` and `minPts`, whose optimal values can be difficult to determine.

This paper introduces an improved implementation of DBSCAN, which reduces the need for specifying both parameters by requiring only the `epsilon` parameter. This enhancement simplifies the application of DBSCAN across different datasets and improves its usability.

The paper is organized as follows: Section 2 explains the fundamental principles of the DBSCAN algorithm. Section 3 discusses various DBSCAN modifications. Section 4 describes our enhanced approach. Section 5 presents the results of applying our method to several benchmark datasets. Finally, Section 6 provides conclusions and suggests future research opportunities.

2. DBSCAN Algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that can detect clusters with arbitrary shapes and sizes. Unlike k-means, it does not require specifying the number of clusters in advance and can identify clusters that are not necessarily spherical. DBSCAN uses two parameters, `minPts` and `eps`, to define density in the data and can classify points as core points, border points, or noise points, making it robust to handle noise. The DBSCAN algorithm is shown below:

Algorithm 1: DBSCAN algorithm

Input

- X: Dataset
- minPts: Minimum number of points
- Eps: Maximum distance

Output

- Labels

1. Identify sets of points within the epsilon neighborhood for each point in X.
2. Choose the first unclassified point:
 - a. If the number of neighbors is less than minPts, label the point as noise and return to step 2.
 - b. Otherwise, mark the current point as a core point of cluster C and add all its neighbors to a queue.
3. Iterate over all unclassified points in the queue:
 - a. If the current point is labeled as noise or has fewer neighbors than minPts, mark it as a border point of cluster C.
 - b. Otherwise, mark the current point as a core point of cluster C and add all its neighbors to the queue.
4. Select the next unclassified point in X and increment the cluster count by 1.
5. Repeat steps 2–4 until all points in X have been labeled.

To assign a point to a cluster, the point must either have at least `minPts` neighbors within its epsilon neighborhood (`eps`), classifying it as a core point, or be within the epsilon neighborhood of a core point, classifying it as a border point. Points that do not meet these criteria are considered noise.

The parameter `minPts` is crucial, as it significantly affects clustering results. If `minPts` is too low, the algorithm may identify too many small clusters or treat noise as clusters. If it is too high, smaller meaningful clusters may be labeled as noise. This sensitivity can make achieving optimal clustering challenging, especially with datasets of varying densities.

3. Related Work

Recent research aimed to simplify DBSCAN by using a single parameter to reduce the complexity of parameter tuning while maintaining high-quality clustering results. Various approaches have been proposed, attracting attention from the research community. This section reviews and discusses these recent papers and their potential to achieve high-quality clustering results with a single parameter.

Yihong et al. [12] propose GNN-DBSCAN, an algorithm using an adaptive grid and dynamic radius based on k-nearest neighbors to overcome DBSCAN's limitations. By defining local and global core samples, GNN-DBSCAN adapts to datasets with large density variations, enhancing core sample quality and clustering performance.

Lu et al. [13] introduce SAG-DBSCAN, a self-adaptive grey DBSCAN clustering algorithm that utilizes a grey relational matrix to obtain the grey local density indicator for adaptive noise

identification. This method automatically estimates clustering parameters and outperforms other state-of-the-art algorithms on several commonly used datasets.

Hu et al. [14] present KR-DBSCAN, a density-based clustering algorithm leveraging reverse nearest neighbor and influence space. This method distinguishes adjacent clusters with varying densities, reduces computational load, and identifies boundary and noise objects. KR-DBSCAN uses reverse nearest neighborhood and influence space for cluster expansion and breadth-first traversal for adding core objects, enhancing clustering accuracy. However, it is more complex and computationally expensive compared to standard DBSCAN, especially for large datasets.

Mu et al. [15] address the limitations of DBSCAN. They propose a parameter-free algorithm combining KNN and genetic algorithms to automatically determine Eps and minPts, improving DBSCAN's accuracy and noise detection. Their experiments on classical datasets and taxi trip records demonstrate the method's superior performance and efficiency compared to the original DBSCAN, validating its ability to handle variable density levels effectively.

The reviewed research suggests that a single-parameter approach to DBSCAN could be a viable alternative to the traditional two-parameter method. Various methods, each with unique strengths and limitations, have been proposed. In the following section, we will present our novel approach that relies solely on a single parameter, eps.

4. Proposed Algorithm

The proposed algorithm is an adaptation of the DBSCAN algorithm that leverages a density-based approach for clustering, utilizing only the epsilon (eps) parameter to define neighborhoods. This algorithm identifies clusters of arbitrary shapes and sizes without the need to pre-specify the number of clusters.

The proposed algorithm begins by computing the neighbors for each point in the dataset within a specified radius (eps). This radius defines the neighborhood around each point. Once the neighbors are identified, the points are sorted in descending order based on the number of neighbors they have. This sorting ensures that the algorithm starts expanding clusters from the densest regions of the dataset. The clustering process is initiated with a cluster ID set to 1. The algorithm then iterates over each point in the sorted list. If a point is unclassified (its label is 0), the algorithm calls the `expandCluster` function to begin cluster expansion from that point. This function assigns the current cluster ID to the point and recursively explores all connected neighbors, labeling them with the same cluster ID. After completing the expansion from a starting point, the cluster ID is incremented, and the process continues until all points are labeled or the size of the cluster is one or two.

Below is the pseudocode for the proposed algorithm:

Algorithm 2: myDBSCAN algorithm

```
Input X: dataset,  
      eps: radius for neighborhood  
Output Labels  
  Compute neighbors for each point within the epsilon radius  
  Sort points by number of neighbors in descending order  
  clusterID = 1  
  for each pointIndex in sortedIndices do  
    if labels[pointIndex] == 0 then  
      expandCluster(pointIndex, clusterID)  
      if currentClusterSize < 3 then  
        Label all unvisited points as noise.  
      break
```

```
        end if
        clusterID = clusterID + 1
    end if
end for

function expandCluster(pointIndex, clusterID)
    if visited[pointIndex] is False then
        visited[pointIndex] = True
        if labels[pointIndex] == 0 then
            labels[pointIndex] = clusterID
        end if
        for each neighborIndex in neighbors[pointIndex] do
            expandCluster(neighborIndex, clusterID)
        end for
    end if
end function
```

This algorithm provides several important benefits. It requires only the eps parameter, simplifying parameter selection and reducing complexity compared to algorithms that need multiple parameters. Its straightforward approach makes it accessible and easy to apply across various datasets. Although the algorithm does not inherently identify clusters with varying densities, it allows for the consideration of small clusters (e.g., of size 1 or 2) as noise, providing flexibility in handling datasets with sparse regions. By treating these small clusters as noise, the algorithm helps ensure that the resulting clusters are meaningful and robust against minor data anomalies. These features make the algorithm a useful tool for clustering applications across different domains.

5. Experimental Evaluation

In the previous section, we introduced an innovative version of the DBSCAN algorithm that uses only the eps parameter for clustering. In this section, we present empirical evidence demonstrating the effectiveness of our new algorithm by conducting a comparative analysis with the original DBSCAN algorithm, which uses two parameters, as well as the k-means algorithm.

To evaluate the effectiveness of our proposed algorithm, we tested it on three synthetic 2-D datasets with varying shapes and densities. The first dataset, called 'Compound', contains 399 points distributed across six clusters, with two clusters intertwined. The second dataset, 'Aggregation', consists of seven clusters with diverse shapes and sizes. The third dataset, 'Complex9', includes 3,031 points divided into nine clusters of different shapes, to which we added some noise points.

Figure 1 demonstrates the clustering results on the first dataset. Our method effectively identified five distinct clusters, designating the sixth as noise, using only the `eps` parameter set to 1.6. In contrast, achieving the same result with the original DBSCAN algorithm required both `eps` and an additional `minPts` parameter set to 5. The k-means algorithm, as expected, struggled to identify non-spherical clusters, especially when one cluster is enveloped by another.

Figure 2 compares the clustering results of the three algorithms. Both our algorithm and the original DBSCAN successfully identified seven distinct clusters; our approach required only one parameter, while DBSCAN needed two. In contrast, the k-means algorithm failed to properly separate two of the clusters.

Figure 3 refers to the 'complex9' dataset, illustrating how both our proposed algorithm and DBSCAN were able to successfully detect nine distinct clusters with `eps = 12` and `minPts = 3`, despite the presence of noisy points that were correctly classified as noise. However, the k-means algorithm, due to its inherent nature, could not handle the noise points effectively.

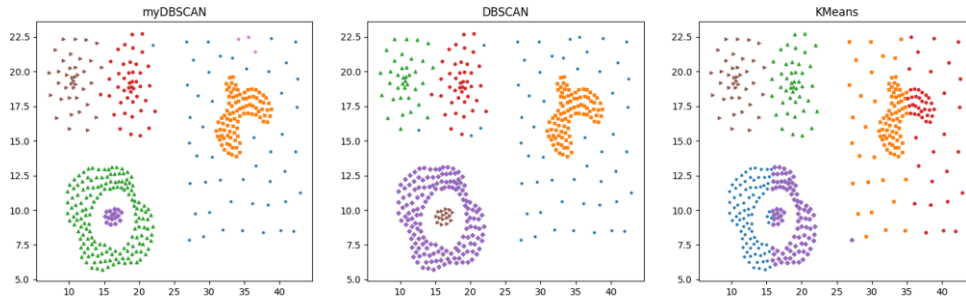


Figure 1. myDBSCAN (a) DBSCAN (b) and k-means (c) clustering results for the Compound dataset.

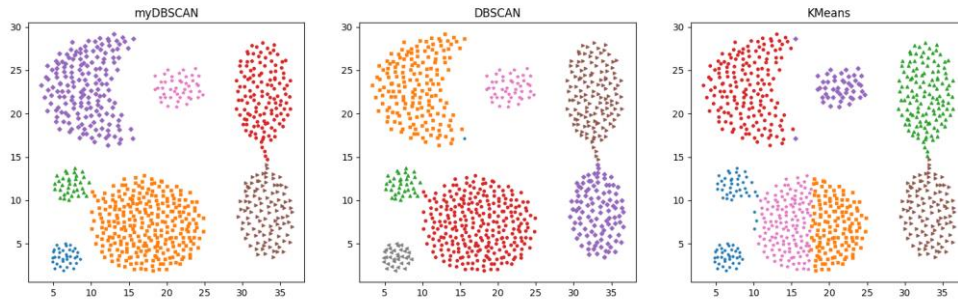


Figure 2. myDBSCAN (a) DBSCAN (b) and k-means (c) clustering results for the Aggregation dataset.

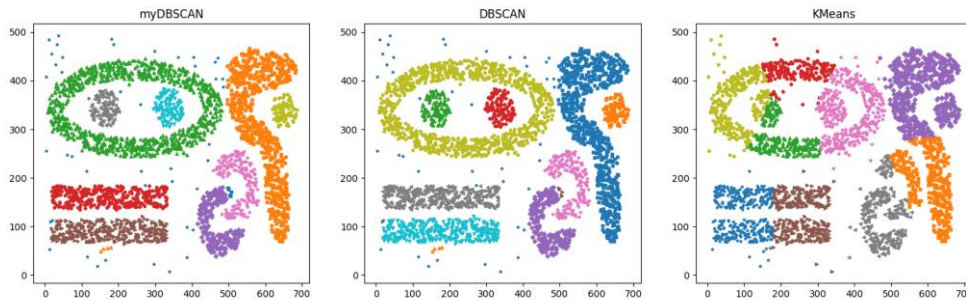


Figure 3. myDBSCAN (a) DBSCAN (b) and k-means (c) clustering results for the complex9 dataset.

To further evaluate the effectiveness of our approach, we conducted experiments on the widely-used Iris dataset. Since this dataset includes true labels, we were able to measure the accuracy of our method and compare it with both the original DBSCAN algorithm and the K-means algorithm. Table 1 provides a summary of the results.

Table1. Performance of Our Method on Real-World Datasets

algorithm	Dataset	iris
	Measure	
myDBSCAN	Accuracy	0.67
	NMI	0.73
	Clu/noise	2/0
	eps	0.1
DBSCAN	Accuracy	0.67
	NMI	0.73
	Clu/noise	2/0

	Minpts/eps	4/.1
k-means	Accuracy	0.97
	NMI	0.90

The table presents the performance of three clustering algorithms—myDBSCAN, DBSCAN, and K-means—on the Iris dataset, evaluated using accuracy and normalized mutual information (NMI). For the myDBSCAN algorithm, the accuracy was 0.67, indicating that 67% of the points were correctly clustered according to the true labels. The NMI score was 0.73, which shows moderate agreement with the true clustering, highlighting its ability to capture some of the structure in the data. MyDBSCAN identified 2 clusters using an epsilon parameter (`eps`) set to 0.1.

Similarly, DBSCAN achieved an accuracy of 0.67, which suggests that both myDBSCAN and DBSCAN perform similarly on this dataset. However, DBSCAN required two parameters—`eps = 0.1` and `minPts = 4`—to achieve these results, highlighting the advantage of myDBSCAN in requiring only one parameter to obtain the same clustering outcome.

In contrast, K-means achieved a significantly higher accuracy of 0.97, demonstrating its effectiveness in clustering the Iris dataset accurately. The algorithm successfully identified 3 clusters, as the number of clusters is provided to the algorithm as a parameter.

Overall, K-means outperforms both myDBSCAN and DBSCAN in terms of accuracy and NMI, indicating its superior ability to handle the Iris dataset. This is likely due to the spherical nature of the clusters in this dataset, which K-means handles well. Both myDBSCAN and DBSCAN detected only 2 clusters, suggesting that the parameter settings might have been too conservative, leading to the merging of clusters that should be separate. The need for setting both `eps` and `minPts` in DBSCAN might have contributed to its suboptimal performance compared to K-means, especially if the true structure involves more defined cluster separations. While DBSCAN-based methods can be useful due to their robustness to noise and ability to find arbitrarily shaped clusters, the Iris dataset's cluster structure appears to favor algorithms like K-means, which assume spherical clusters. Overall, the results highlight that while DBSCAN-based methods can be useful, the simplicity and accuracy of K-means make it better suited for datasets like Iris, where cluster boundaries are more distinct and less noisy.

6. Conclusion

In this study, we introduced a simplified version of the DBSCAN algorithm, referred to as myDBSCAN, which relies on a single parameter, `eps`, for clustering. This modification addresses the common challenge of selecting optimal parameters for clustering algorithms and simplifies the clustering process without sacrificing performance. Our results demonstrate that myDBSCAN is as effective as the original DBSCAN algorithm in identifying clusters in datasets with varying shapes. While both myDBSCAN and DBSCAN produced similar results, with each identifying two clusters, the advantage of myDBSCAN lies in its use of only one parameter compared to the two required by DBSCAN. This makes myDBSCAN more user-friendly and less sensitive to parameter selection.

Overall, the results suggest that myDBSCAN is a viable alternative to traditional DBSCAN, offering simplicity and efficiency in clustering tasks. Future work could focus on further refining the algorithm and exploring its applicability to other datasets with complex structures and varying noise levels.

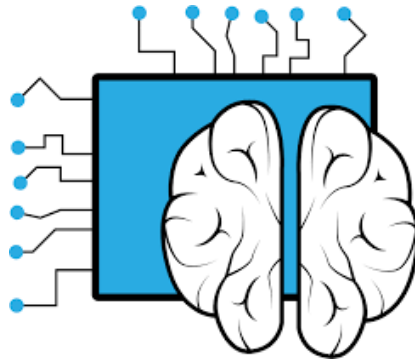
References

- [1] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Transactions on neural networks, vol. 16, p. 645–678, 2005.
- [2] H. Yu and X. Hou, "Hierarchical clustering in astronomy," Astronomy and Computing, vol. 41, p. 100662, 2022.

- [3] R. Nugent and M. Meila, "An overview of clustering applied to molecular biology," *Statistical methods in molecular biology*, p. 369–404, 2010.
- [4] R. C. T. Lee, "Clustering analysis and its applications," in *Advances in Information Systems Science: Volume 8*, Springer, 1981, p. 169–292.
- [5] S. Abolmakarem, F. Abdi and K. Khalili-Damghani, "Insurance customer segmentation using clustering approach," *International Journal of Knowledge Engineering and Data Mining*, vol. 4, p. 18–39, 2016.
- [6] S. Khanmohammadi, N. Adibeig and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," *Expert Systems with Applications*, vol. 67, p. 12–18, 2017.
- [7] M. Ahmed, R. Seraj and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, p. 1295, 2020.
- [8] R. Sibson, "SLINK: an optimally efficient algorithm for the single-link cluster method," *The computer journal*, vol. 16, p. 30–34, 1973.
- [9] M. Ouyang, W. J. Welsh and P. Georgopoulos, "Gaussian mixture clustering and imputation of microarray data," *Bioinformatics*, vol. 20, p. 917–923, 2004.
- [10] M. Ester, H.-P. Kriegel, J. Sander, X. Xu and others, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, 1996.
- [11] M. Ankerst, M. M. Breunig, H.-P. Kriegel and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM Sigmod record*, vol. 28, p. 49–60, 1999.
- [12] L. Yihong, W. Yunpeng, L. Tao, L. Xiaolong and S. Han, "GNN-DBSCAN: A new density-based algorithm using grid and the nearest neighbor," *Journal of Intelligent & Fuzzy Systems*, vol. 41, p. 7589–7601, 2021.
- [13] S. Lu, L. Cheng, Z. Lu, Q. Huang and B. A. Khan, "A Self-Adaptive Grey DBSCAN Clustering Method.," *Journal of Grey System*, vol. 34, 2022.
- [14] L. Hu, H. Liu, J. Zhang and A. Liu, "KR-DBSCAN: A density-based clustering algorithm based on reverse nearest neighbor and influence space," *Expert Systems with Applications*, vol. 186, p. 115763, 2021.
- [15] B. Mu, M. Dai and S. Yuan, "DBSCAN-KNN-GA: a multi Density-Level Parameter-Free clustering algorithm," in *IOP Conference series: materials science and engineering*, 2020.

SESSION 4

Intelligent Systems and Automation



Modernizing Assessment: Overview of Computerized Adaptive Testing and AI

Haddad Dhai Eddine^{1□}, Khentout Chabane¹, Harbouche Khadidja¹

¹ University Setif 1 Ferhat Abbas UFAS1, Computer science Department, LRSD laboratory, Setif,
Algeria

dhaieddine.haddad@univ-setif.dz

Abstract

This paper presents a comprehensive overview of Computerized Adaptive Testing systems, introducing educational field practitioners into the topic of assessment theory. The mechanism to which Computerized Adaptive Testing systems administer tests to learners are explained; exploring how they item response theory is used. The paper further investigates the integration of Machine Learning models to enhance the assessment process. Furthermore, the study highlights both learners and teachers reactions to these new solutions.

Keywords: Computerized Adaptive Testing, Item Response Theory, Machine Learning

1. Introduction

The Covid-19 pandemic, declared by the WHO in early 2020 [1], forced universities worldwide to transition from traditional, in-person learning and assessment to online alternatives. This shift sparked a surge in interest for computerized assessment methods, leading to a greater focus on Computer-ized Adaptive Testing (CAT).

CAT systems, as described by [2], offer a personalized testing experience by adapting the difficulty level of questions based on the student's performance. This approach allows for more accurate ability estimates, quicker results, and shorter testing times compared to traditional methods. The core of a CAT system lies in two key components: Item Response Theory (IRT) and item selection algorithms. IRT models the relationship between student ability and the probability of answering a question correctly. Item selection algorithms, guided by IRT, choose the next item based on the student's previous responses, gradually increasing or decreasing difficulty to identify their true ability level.

This paper aims to serve three main purposes. First, we explore both traditional and modern assessment methodologies, explaining key terminology and defining fundamental concepts, introducing the reader to the realm of assessments. Second, we delve into the workings of CAT systems, focusing on IRT and item selection algorithms, to define these systems and their core components.

Finally, we examine the integration of Machine Learning (ML) and Deep Learning (DL) solutions in modern CAT systems, focusing on recent advancements in CAT utilization.

2. ASSESSMENT

In the field of educational measurement, assessment plays a vital role in understanding human learning and ability. This is usually done through the analysis of responses to standardized tasks (e.g. Algorithms and Data structures test for CS students). The field has always relied on Classical Test Theory (CTT), which is considered the standard across various educational settings and around the world.

However, with the recent advances in both computing powers and technology being rooted in the lifestyles of new generation of students, Item Response Theory has emerged as an alternative approach, addressing the limitations of CTT.

3. CAT SYSTEM COMPONENTS

According to the literature [3] [4] [5], Computerized Adaptive Testing systems are made of the following components:

3.1 IRT model

Item Response Theory (IRT) is a set of statistical models that can be used for test construction, administration and analysis [6]. Unlike Classical test theory, it shifts the focus from total scores to questions which makes it highly adaptable as well as rendering it the corner stone for many well renowned international large scale assessment programs such as PISA, OECD, NAEP and china's National Assessment of Basic Education Quality [7]. IRT utilizes multiple statistical models to analyze tests and provide a more precise understanding of students abilities.

3.1.1 One parameter logistic model (1PL)

Also known as the Rasch model after its creator. This is the simplest IRT model and can be described with equation (1) mentioned bellow

$$P(x = 1|\theta) = \frac{e^{(\theta - \beta)}}{1 + e^{D(\theta - \beta)}} \quad \text{Eq.1}$$

where $P(X=1)$ means the probability of correctly answering an item given ability level θ . β represents the difficulty parameter. and lastly D is a scaling constant (equal to 1.7).

3.1.2 Two parameters logistic model (2PL)

It builds upon the 1PL model by introducing an additional item parameter called the "discrimination" parameter " α ". This parameter one of, if not the most important aspect on an item as it helps teachers distinguish between good and bad students. Test developers always seek to incorporate discriminating items into their tests further increasing the importance of this parameter. The model is defined with equation (2):

$$P(x = 1|\theta) = \frac{e^{\alpha(\theta - \beta)}}{1 + e^{\alpha D(\theta - \beta)}} \quad \text{Eq.2}$$

3.1.3 Three parameters logistic model (3PL)

The 3PL model further expands on the 2PL by introducing a third item parameter; the pseudo-guessing parameter C . Given a multiple choice item, even a student with low ability θ has a chance of correctly answering it. This possibility can be noticed looking at the Item Characteristic Curve (ICC) where the probability of a correct answer never reaches zero, see fig. 1 for reference.

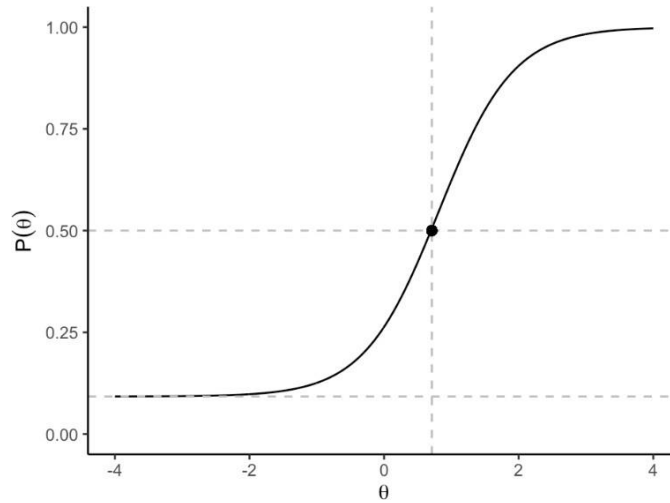


Figure 1. Item characteristics curve 3PL

$$P(x = 1|\theta) = C + (1 - C) \frac{e^{\alpha(\theta - \beta)}}{1 + e^{\alpha(\theta - \beta)}} \quad \text{Eq.3}$$

3.1.4 Four parameters logistic model (4PL)

a relatively underutilized and underappreciated model due to its increased complexity, the 4PL model [8] introduces yet another parameter called the upper asymptote parameter δ . Conversely to 3PL, this model accounts for the blunders of good students, acknowledging the fact that even top students make mistakes. Due to CAT tests sequential nature and the inability to revise answers, this model can help increase ability estimation precision as it allows for room for error [9].

$$P(x = 1|\theta) = C + (\delta - C) \frac{e^{\alpha(\theta - \beta)}}{1 + e^{\alpha(\theta - \beta)}} \quad \text{Eq.4}$$

It should be noted that while logistic models are more common in the literature, there exists "nominal" models in IRT too. This paper will focus on logistic models for their popularity, for more information on the nominal IRT models see [10]. Another point worth mentioning, the 4PL model is the general form and 1PL, 2PL, 3PL can be regarded as special cases (e.g. 1PL model is a special case of 4PL with the upper asymptote parameter δ set to 0, the pseudo-guessing parameter C set to 0 and the discrimination parameter α set to 1).

3.2 Item bank

Item pools or banks, as stated by [11], are the cornerstone of CAT systems. They are a curated collection of items representing a specific ability or skill, such as programming or math. The word "calibrated" means the psychometric properties of these items such as difficulty, discrimination etc are estimated. These properties are also part of the information included in the item bank [12]. To highlight the importance of this component, we take the example of a newly appointed teacher. He lacks the expertise to craft good tests which leads to inaccurate ability estimates. This can be solved by using an item bank [13] which consists of well-designed items with known properties making it possible for the teacher to get good ability estimates and even target specific student populations.

3.3 Starting point

In general, the systems pick a medium difficulty item. Even if the student possesses a sub-average ability level and the question appears hard for him, it wouldn't affect the overall test as the CAT system will eventually adjust to his ability level [5].

3.4 Item Selection Algorithms

3.4.1 Maximum Fisher Information (MFI)

This algorithm takes a simple approach, selecting the item at a given θ level that maximizes Fisher's information function [14].

3.4.2 α stratification

α -stratification improves item exposure rates. It divides the item pool into strata based on increasing α parameter values.

3.4.3 Global information

Another approach designed to address limitations of the MFI method. Introduced by [15], this algorithm utilizes the moving average of Kullback-Leibler information to strategically select items during a test. This algorithm improves usage rate of the item pool.

3.4.4 Likelihood weighted information

Introduced by [16], incorporates the likelihood function to weight the Fisher information function. This weighted information is then summed across the entire ability range (θ) after the administration of n items.

The landscape of item selection algorithms for computerized adaptive testing (CAT) offers a diverse range of options. Some, like the Maximum Fisher Information (MFI) method, are known for their simplicity and intuitive nature. Others, like the α -stratification and Global Information approaches, were specifically developed to address limitations of earlier methods. Additionally, some algorithms excel at providing informative assessments across the entire ability range (θ), while others may be more effective at the extreme such as the Likelihood weighted information (LWI) algorithm.

3.3 Termination Rule (criterion)

Computerized Adaptive Testing (CAT) utilizes various termination rules to determine when to end the test for each examinee. These rules can be broadly categorized into three types [4]: ability-based, focusing on the estimated ability level (θ) reaching a desired level of precision; error-based, stopping when the standard error of measurement falls below a pre-defined threshold; and item bank usage

rules, ensuring items are not overused and maintaining fairness. The optimal choice of termination rule depends on the specific assessment's goals and context, balancing factors like accuracy, efficiency, and item bank preservation [12].

3.4 Scoring method

In Computerized Adaptive Testing (CAT), estimating an examinee's ability (θ) is key. Two main approaches exist: Maximum Likelihood Estimation (MLE) offers simplicity and efficiency and produces results that are less biased [17]. Conversely, Bayesian methods tend to have greater bias values but they make up in producing estimates with smaller MSE [12].

4. ARTIFICIAL INTELLIGENCE AND CAT

The surge of Artificial intelligence (AI) research is evident, with researchers actively exploring the use of machine learning (ML) and deep learning (DL) solutions for existing problems. Computerized Adaptive Testing (CAT) is no exception, as research is actively being conducted to integrate these new technologies into existing solutions.

One of the use cases is improving Item Response Theory (IRT) with ML/DL. This area focuses on enhancing the capabilities of IRT models, which are traditionally used in CAT, by incorporating ML/DL techniques. This can lead to faster plus more accurate ability and item parameters estimation. For instance, [18] their findings suggest that this approach led to faster parameter estimation and even allowed for the extension of IRT models using Deep Learning (DL) to make use of large datasets provided by E-Learning platforms. Additionally, research by [19] combined IRT with Dynamic Key-Value Memory Networks (DKVMN). the latter estimates ability θ and difficulty parameter β and passes it to IRT model to make predictions.

A noteworthy use case is presented in the work of [20]. By harnessing the capabilities of Reinforcement Learning, they achieved the development of a novel item selection algorithm. Similarly, [21] approaches the problem of item selection through a bi-level optimization lens, resulting in a data-driven, model-agnostic algorithm that surpasses previously employed, handcrafted solutions. it is crucial to acknowledge that these methods come with their own set of challenges. Firstly, they often require vast amounts of data, which may not always be readily available or pose ethical concerns regarding data collection and usage. Secondly, the very nature of data-driven approaches can introduce potential biases, necessitating careful consideration and mitigation strategies to ensure fairness and accuracy.

Beyond the previously mentioned approaches, another promising solution involving Machine Learning (ML) is the combination of Decision Trees (DT) with Computerized Adaptive Testing (CAT). This approach, as explored in [22], allows for the creation of CATs that not only minimize the mean squared error in estimating examinee ability levels but also control the exposure rate of individual items. [23], similar to the previous method, it tries to mitigate its computational limitations by making test generation feasible on a personal computer, thereby enhancing accessibility and potential applications.

5 Datasets

Recognizing data as the cornerstone of any research, we have compiled information on datasets utilized in the literature of assessment. This resource aims to assist new researchers by simplifying their search for high-quality data, ultimately speeding up their research. Additionally, teachers with existing digital past assessments can leverage this resource without relying on the specific datasets provided. Generating synthetic response data is another viable option. Tools like Simulcat and mIRT packages facilitate this process, making it a common practice in the field.

Table1. DATASETS USED IN CAT RESEARCH

Name	Description	Accessibility
ASSISSTment	real response data	freely available
Khan Academy	real response data	On demand
Datashop	real response data	freely available
KDcup	real response data	freely available

6. Cat reception

While technology is now ubiquitous, gauging the perspectives of both teachers and students on CAT is crucial for a smooth and successful transition from traditional paper-based assessments to digital e-learning solutions. This feedback can ensure a clear, simple, and cost-effective implementation process.

6.1. Teachers perspective

Teacher perspectives are crucial for successful CAT implementation. Studies like [24] indicate that teachers are generally comfortable using the technology. However, a teacher in [24] expressed concerns about limited ability to pre-review the entire test before administration. Despite these reservations, research by [25] suggests a growing trend of teacher adoption. Notably, [26]'s framework highlights teacher interest in CAT but emphasizes the need for support. Teachers in this study expressed a preference for readily available technical assistance over costly infrastructural changes.

6.2 Students Perspective

As students are the primary users of CAT, understanding their interaction and experience with the system is crucial before transitioning from traditional testing. Fortunately, existing research provides valuable insights in this area. For example, a study by [27] applied CAT to surgery students and reported positive user reactions. They perceived CAT as practical and fairer than traditional testing using Classical Test Theory (CTT). Italian health informatics students at the University of L'Aquila Italy [24] found no significant difficulties associated with using CAT. In fact, the results indicated that students spent less time answering individual questions compared to CTT. This aligns with existing research suggesting that CAT can substantially reduce test length. [25] highlights several noteworthy findings regarding CAT.

Notably, the system was well-received by both older students and even those who failed in the initial assessment. These students perceived their results as accurate and fair.

7. Conclusion

This paper offers a comprehensive exploration of Computerized Adaptive Testing (CAT), delving into its core components, theoretical foundations, and implementation using cutting-edge AI techniques. Recognizing the evolving landscape of CAT, it also discusses current trends and applications, while acknowledging the crucial perspectives of both students and teachers in facilitating a seamless transition from conventional testing methods to these state-of-the-art solutions.

References

- [1] M. Ciotti, M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, and S. Bernardini, "The COVID-19 pandemic," *Critical Reviews in Clinical Laboratory Sciences*, vol. 57, no. 6, pp. 365–388, Aug. 2020, doi:10.1080/10408363.2020.1783198
- [2] H.-H. Chang, "Psychometrics Behind Computerized Adaptive Testing," *Psychometrika*, vol. 80, no. 1, pp. 1–20, Mar. 2015, doi: 10.1007/s11336-014-9401-5
- [3] N. Thompson, "A Practitioner's Guide for Variable-length Computerized Classification Testing," *Practical Assessment, Research and Evaluation*, vol. 12, Jan. 2007
- [4] N. A. Thompson and D. A. Weiss, "A Framework for the Development Of Computerized Adaptive Tests," *Practical Assessment, Research, and Evaluation*, vol. 16, no. 1, Art. no. 1, Nov. 2019, doi: 10.7275/wqzt-9427
- [5] D. J. Weiss and G. G. Kingsbury, "Application of Computerized Adaptive Testing to Educational Problems," *Journal of Educational Measurement*, vol. 21, no. 4, pp. 361–375, 1984
- [6] S. P. Reise, A. T. Ainsworth, and M. G. Haviland, "Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research," *Current Directions in Psychological Science*, vol. 14, no.2, pp. 95–101, 2005
- [7] L. Cai, K. Choi, M. Hansen, and L. Harrell, "Item Response Theory," *Annu. Rev. Stat. Appl.*, vol. 3, no. 1, pp. 297–321, Jun. 2016, doi:10.1146/annurev-statistics-041715-033702
- [8] K. L. Rulison and E. Loken, "I've Fallen and I Can't Get Up: Can High-Ability Students Recover From Early Mistakes in CAT?," *Applied Psychological Measurement*, vol. 33, no. 2, pp. 83–101, Mar. 2009, doi:10.1177/0146621608324023
- [9] M. A. Barton and F. M. Lord, "An Upper Asymptote for the Three-Parameter Logistic Item-Response Model*," *ETS Research Report Series*, vol. 1981, no. 1, pp. i–8, 1981, doi: 10.1002/j.23338504.1981.tb01255.x
- [10] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887, 1986
- [11] D. Cella, R. Gershon, J.-S. Lai, and S. Choi, "The Future of Outcomes Measurement: Item Banking, Tailored Short-Forms, and Computerized Adaptive Assessment," *Quality of Life Research*, vol. 16, pp. 133–141, 2007
- [12] D. O. Segall, "Computerized Adaptive Testing," in *Encyclopedia of Social Measurement*, Elsevier, 2005, pp. 429–438. doi: 10.1016/B0-12-369398-5/00444-8
- [13] Choppin, "Item Bank using Sample-free Calibration," *Nature*, vol. 219, no. 5156, pp. 870–872, Aug. 1968, doi: 10.1038/219870a0
- [14] D. J. Weiss, "Improving Measurement Quality and Efficiency with Adaptive Testing," *Applied Psychological Measurement*, vol. 6, no. 4, pp. 473–492, Sep. 1982, doi: 10.1177/014662168200600408

- [15] H.-H. Chang and Z. Ying, "A Global Information Approach to Computerized Adaptive Testing," *Applied Psychological Measurement*, vol. 20, no. 3, pp. 213–229, Sep. 1996, doi:10.1177/014662169602000303
- [16] W. J. J. Veerkamp and M. P. F. Berger, "Some New Item Selection Criteria for Adaptive Testing," *Journal of Educational and Behavioral Statistics*, vol. 22, no. 2, pp. 203–226, 1997, doi:10.2307/1165378
- [17] M. Wu, R. L. Davis, B. W. Domingue, C. Piech, and N. Goodman, "Variational Item Response Theory: Fast, Accurate, and Expressive," 2020, doi: 10.48550/ARXIV.2002.00276
- [18] C.-K. Yeung, "Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory." *arXiv*, Apr. 26, 2019, doi: 10.48550/arXiv.1904.11738
- [19] Y. Zhuang, Q. Liu, Z. Huang, Z. Li, S. Shen, and H. Ma, "Fully Adaptive Framework: Neural Computerized Adaptive Testing for Online Education," *AAAI*, vol. 36, no. 4, pp. 4734–4742, Jun. 2022, doi:10.1609/aaai.v36i4.20399
- [20] A. Ghosh and A. Lan, "BOBCAT: Bilevel Optimization-Based Computerized Adaptive Testing." *arXiv*, Aug. 16, 2021. [Online]. Available: <http://arxiv.org/abs/2108.07386>
- [21] D. Delgado-Gómez, J. C. Laria, and D. Ruiz-Hernandez, "Computerized adaptive test and decision trees: A unifying approach," *Expert Systems with Applications*, vol. 117, pp. 358–366, Mar. 2019
- [22] J. Rodriguez-Cuadrado, D. Delgado-Gómez, J. C. Laria, and S.
- [23] Rodriguez-Cuadrado, "Merged Tree-CAT: A fast method for building precise computerized adaptive tests based on decision trees," *Expert Systems with Applications*, vol. 143, p. 113066, Apr. 2020, doi:10.1016/j.eswa.2019.113066
- [24] A. M. Angelone, A. Galassi, and P. Vittorini, "Lessons learned about the application of adaptive testing in several first-year university courses," *Int. J. Learn. Technol.*, vol. 17, no. 1, pp. 3–26, Jan. 2022, doi: 10.1504/ijlt.2022.123696.
- [25] C. F. Collares and D. Cecilio-Fernandes, "When I say . . . computerised adaptive testing," *Medical Education*, vol. 53, no. 2, pp. 115–116, Feb. 2019, doi: 10.1111/medu.13648
- [26] A. Fink, C. Spoden, and A. Frey, "Determinants of higher education teachers' intention to use technology-based exams," *Educ Inf Technol*, vol. 28, no. 6, pp. 6485–6513, Jun. 2023, doi: 10.1007/s10639-022-11435-4
- [27] A. M. Angelone and P. Vittorini, "A Case Study on Students' Opinions About Adaptive and Classical Tests," in *Methodologies and Intelligent Systems for Technology Enhanced Learning*, 12th International Conference, M. Temperini, V. Scarano, I. Marenzi, M. Kravcik, E. Popescu, R. Lanzilotti, R. Gennari, F. De La Prieta, T. Di Mascio, and P. Vittorini, Eds., in *Lecture Notes in Networks and Systems*. Cham: Springer International Publishing, 2023, pp. 26–36. doi:10.1007/978-3-031-20617-7_5

Deep Reinforcement Learning for Multi-intersection Traffic Signal Control: A New Cooperative Approach

Tarek Amine Haddad

*LaSTIC laboratory, Department of Computer Science, University of Batna 2, Batna 05078 Algeria.
t.haddad@univ-batna2.dz*

Abstract

Traffic congestion is a significant cause of economic losses and in-convenience in urban areas. Recently, the widespread use of traffic detectors and advancements in data processing algorithms has enabled a data-driven approach to alleviating congestion. Deep reinforcement learning (DRL) offers a promising approach for achieving intelligent traffic light control (TSC). However, managing multiple intersections poses challenges, mainly because of the requirement to ensure effective communication among expanding intersections and spatiotemporal dynamics play a significant role in traffic flows. This paper proposes a novel cooperative approach based on DRL and pheromone-based state representation. This approach allows intersections to exchange information and promote collaboration by receiving information from neighboring intersections in multi-intersection situations, thus improving cooperation and reaching decisions that are globally optimal. Experimental findings in multi-intersection traffic scenarios show that the proposed approach exhibits performance benefits across a variety of metrics, validating the efficacy of the proposed approach.

Keywords: traffic congestion, traffic signal control, deep reinforcement learning, multi-intersection, pheromone.

1. Introduction

Urban environments are seriously affected by traffic congestion, because of its significant financial burden and the adverse effects on the area's residents, resulting from air pollution, stress-induced nervous conditions and emergency services are delayed [1]. That is why significant endeavours have been made to address this issue, particularly in urban areas. Consequently, various proposals have emerged for Intelligent Transportation Systems (ITS) to identify congestion and manage it through vehicle re-routing, aiming to minimize waiting time for vehicles and enhance traffic flow. To ease congestion, implementing traffic signal control (TSC) based on traffic conditions has proven to be an efficient strategy [2].

With advancements in artificial intelligence, deep reinforcement learning (DRL) has become increasingly prevalent in ITS, offering a novel potential to enhance safety, sustainability, and efficiency. The algorithm DRL, developed within the framework of Markov Decision Processes (MDP), is known for its better learning ability in complex dynamic environments [3]. In the last decade, TSC field has seen the emergence of numerous advanced approaches applied to isolated intersection based on DRL [4], [5], [6]. However, the objective of reducing traffic congestion presently is frequently not accomplished by solely controlling and optimizing the traffic signals at isolated intersections. On the contrary, it necessitates collaboration among multiple intersections, moving beyond local to global improvement. This involves coordinating the control of multiple intersections

towards achieving a sub-optimal or optimal solution for the entire transportation network, oriented towards TSC.

In a multi-intersection setting, the common approach to signal control is to allocate a separate agent to each intersection. Each agent then manages its respective intersection independently, or cooperative control can be achieved by sharing state information among adjacent agents [7], [8], [9]. In other words, the cooperative TSC for multi-intersections assists in regulating traffic flows at adjacent intersections, enhancing the effectiveness of traffic management, and potentially preventing traffic congestion issues. To this end, some approaches in Multi-Agent Reinforcement Learning (*MARL*) have aimed to create learning frameworks suitable for extensive agent collaboration, which include fully centralized methods [10], [11] or decentralized methods [12], [13]. However, these approaches aim to find the best approach for coordinating multiple agents, necessitating extensive communication among agents and scalability.

To achieve cooperative *TSC*, we need traffic information from nearby intersections and on-road vehicles. In this paper, a *DRL*-based cooperative approach to address the issue of *TSC* at multiple intersections. In this scenario, each traffic signal at an intersection is treated as a separate agent responsible for learning decision-making processes that take into account the traffic conditions of both local and neighbouring intersections. The pheromone-based state representation effectively extracts information. To demonstrate the advantages of the proposed approach, we conduct an extensive simulation study, including a comparison with traditional methods.

The rest of this manuscript is organized as follows: Section 2 offers an overview of related work. The problem is presented in Section 3. Experiments are performed in section 4 and the experimental results are presented in section 5. Finally, draw some conclusions in section 6.

2. Related Works

Reinforcement learning is a goal-driven approach that relies on interactions between the environment and agent. Nonlinear approximation and hierarchical feature extraction are capabilities in which deep learning (*DL*) excels. The combination of these methods, known as *DRL*, offers many attractive features. Consequently, the interest of many researchers has been captured by applying *DRL* to *ATSC*.

In recent years, there has been extensive research on *DRL* algorithms within the realm of *TSC* [14]. The *TSC* process is regarded as a *MDP*, utilizing *DRL* methods to generate traffic signal policies. In research [15], neural networks are employed to enact real-time *TSC*. Deep neural networks can enhance the effectiveness of RL significantly. Hence, a variety of *DRL* algorithms, including deep Q-network (*DQN*) [16], are utilized for implementing *TSC*. Numerous RL-based methods focus on single intersection. Multiple intersections are involved in city road networks. Numerous approaches will encounter the challenge of the curse of dimensionality when dealing with environments involving multiple intersections [17], [8]. Chen et al. [18] tackle the issue of multi-intersection based on *RL* methods. They address issues of signal coordination, scalability, and data feasibility by applying pressure to ensure the coordination of signals. In [19] authors adopt method with each agent learns dynamically and independently adjusts its policy. The research [20] develops various agents tasked with controlling diverse shape intersections. A cooperative approach coordinates several agents to generate collaborative *TSC* is proposed in [21], in this approach, each intersection is interacted with by each agent, as well as adjacent agents.

In [22], the authors propose a cooperative *DRL* framework to address the problem of reducing traffic congestion. It comprises centralized global agent and several regional agents. The value function and policy of each regional agent are learned independently within its small region. The centralized agent

consolidates the *RL* outcomes from various regional agents across the extensive traffic grid in a hierarchical manner.

As reviewed above, in many methods that encourage collaboration, information is merged or shared directly without considering its importance. Thus in this work, a new MARL-based cooperative model is designed to reduce traffic congestion in multi-intersection. Based on pheromone-information and *Double DQN* method, the designed model optimizes the entire algorithm from various perspectives simultaneously. That means, additionally, this work can reduce four important metrics such as Average Waiting Time (AWT), Average Emission CO2 (AEC), Average Pheromone Intensity (API), Average Noise Emission (ANE).

3. Problem Definition

In this section, the appropriate definitions of intersections are introduced. The traffic environment is regarded as a grid of multiple intersections in this work. In the model proposed, each intersection is controlled by an independent agent. The movement is regulated by traffic signals. The indices set i ($i = [1, 2, \dots, I]$) denotes both intersections and agents. I represent the total number of intersections. As depicted in Fig. 1, a typical cross intersection comprises four directions, labeled *West* (W), *South* (S), *East* (E), and *North* (N) based on their orientations. Intersection i is defined as a junction where each entry consists of lanes and each exit is equipped with lanes. The exits and entrances denote the lanes at the intersection that vehicles specifically use to enter or exit the road. The state comprises two components: the intersection's individual state and the road network's global state are being considered. a_i^t is the action carried out by the agent. In the TSC task context, actions denote the modifications the agent applies to the phase of lights. To ensure road traffic safety, vehicle flows are kept separate and do not intersect. Typically, action settings encompass four different phases as shown in Fig.2. In the phase sequence definition, each phase is executed in turn when it is green, while the other phases display a red signal during the same cycle.

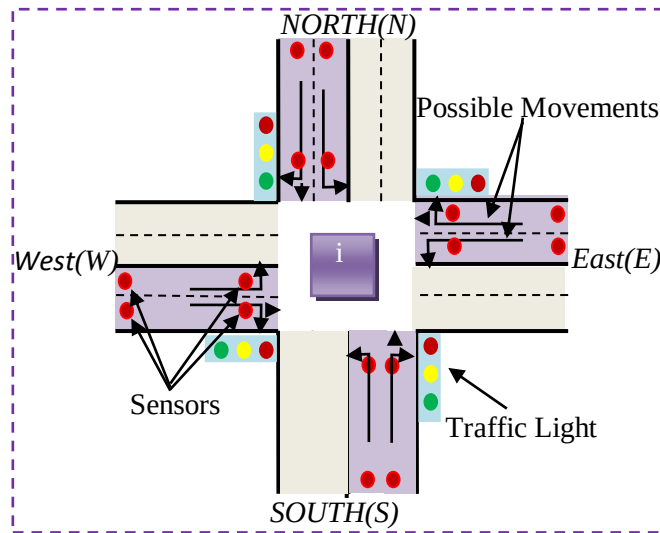


Figure 1. Intersection definitions with traffic movements

As previously noted, within the road network, the traffic situation at an intersection is determined by the signal control within the intersection itself, as well as by the traffic flows of neighboring intersections. Hence, a cooperative *TSC* approach that takes into account the traffic conditions of these interconnected intersections is crucial for achieving optimal signal control across the entire system. This study addresses the issue of cooperative *TSC* within a network of roads featuring multiple signalized intersections. The primary goal of this study is to alleviate traffic congestion at multi-intersection.

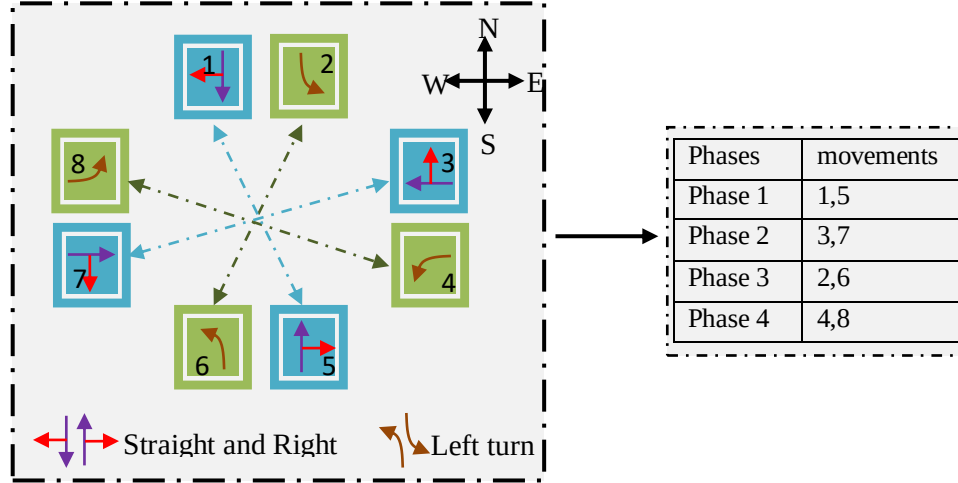


Figure 2. Traffic signal phases design

4. Proposed Multi-Agent Traffic Signal Control Approach

A cooperative control model for multiples intersections using a *MARL* method is introduced in this paper. The framework is illustrated in Fig. 3. Both the local state s_i of each agent and its neighbor are taken as input of the network. The output is the collective *Q-value* that includes all agents. The primary environmental attributes include intersection phases, vehicle speeds, and the count of cars at specific intersections. More attributes may be incorporated into the state, such as lane capacities, intersection capacities, and entry or exit speeds at the intersection. In this paper, we use queue length Q , pheromone intensity ph and action a_i as a state of each intersection.

The primary goal of the proposed system is to reduce the total waiting time for vehicles at traffic intersections. The reward r_i is expressed as that of an agent i as follow:

$$r_i = \begin{cases} 1 & \text{if } f_{t+1}^i > f_t^i \\ -1 & \text{if } f_{t+1}^i < f_t^i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where f represents the total traffic flow at intersection i . The reward function's design should steer the agent towards achieving better training outcomes, particularly in easing traffic congestion. Therefore, the reward function of each agent must take into account both the traffic flow at the intersection and the traffic flow of its neighbors to minimize overall traffic congestion.

This paper utilizes the *DDQN* algorithm, an enhancement over the original *DQN*. To stabilize the *DDQN* algorithm, the target network parameters are set to be fixed after a designated number of steps. The formula is defined as:

$$Q_{target} = r_t + \gamma Q_t(s_{t+1}, \arg \max_a Q_t(s_{t+1}, a', \theta_i); \theta') \quad (2)$$

Where r_t is the reward, θ_i is the evaluation network and γ is the discount factor. In the *DDQN* method, the action is determined not by the older *Q-network*, but by selecting the action with the highest *Q-value* from the current *Q-network*. As a result, the problem of *Q-value* overestimation is addressed.

The ϵ -greedy policy [23] is used for action selection, wherein actions are selected based on their *Q-values*, with the action having the highest *Q-value* chosen with a probability of $1 - \epsilon$, and with a probability of ϵ , a random action is selected from the action set.

In this work, the proposed framework consists primarily of an agent and a traffic environment. The traffic environment comprises a network of roads with grid of 2×2 intersections. Each agent oversees a

single intersection, where its actions directly influence the management and phase for that intersection in the subsequent cycle. The agents have the capability to exchange traffic information with their neighboring agents.

In addition to focusing on the local state and reward for each agent should also include the state and reward information of adjacent agents when calculating its own Q value, and ultimately achieve the global optimal action. Fig. 3 illustrates the interactions between the agent and the traffic environment at time step t .

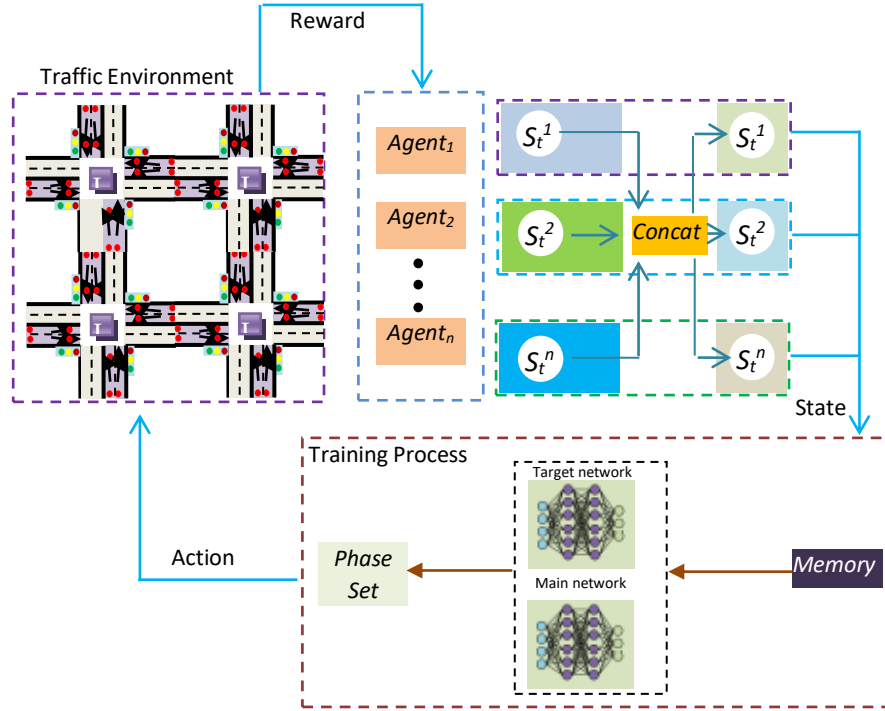


Figure 3. The framework of proposed approach

Our approach consistently chooses the optimal action, which corresponds to the correct traffic light phase at multi-intersection, aiming to reduce four metrics: AWT, AEC, API and ANE.

During the training process, the policy of each agent, defined by its Q-function, is estimated using a category of functions commonly referred to in research as deep neural networks (DNN). The state in vector form is fed into an input layer of each DNN, and produces a vector as output, where each element matches to the assessment of the action.

The pseudocode of the proposed Algorithm is presented below. At each step t , each agent gets state of intersection s_t and its neighbors from the road network. Afterward, the action a_t is chosen and carried out by each agent using a greedy strategy. Fig. 2 comprehensively illustrates all possible actions. Following this, the environment transitions to the subsequent state s_{t+1} and delivers the corresponding reward r_t . Therefore, they are all stored in memory as tuples (s_t, a_t, r_t, s_{t+1}) . Once the memory accumulates a certain number of samples, it used a randomly selected set of these experiences to adjust the weights of neural network. By utilizing the *Adam* (Adaptive Moment Estimation) optimizer to perform backpropagation for training DNN, while minimizing loss function. Additionally, it introduces the concept of a target network, significantly enhancing the learning process. Finally, Algorithm 1 produces the Q-function, which determines the behavior of the agent.

Algorithm 1: MARL Algorithm for multi-intersection TSC

```

1: Input : episodes  $Z$ , discount factor  $\gamma$ , Memory size  $M$ 
2: Output : parameters  $\theta$ 
3: Initialization:  $Z$ , total steps  $T$ ,  $\theta$ ,  $\theta'$ 
4: Begin
5:   for episode = 1 to  $Z$  do
6:     Initialize the environment
7:     for  $t = 1$  to  $T$  do
8:        $s_i(t) = \text{get individual state}$ 
9:        $a_i(t) = \text{select action according to the } \varepsilon\text{-greedy}$ 
10:      execute  $a_i$ 
11:      get reward  $r_i(t)$  and next state  $s_{t+1}$  in Eq.(1)
12:      Add  $(s_i, a_i, r_i, s_{t+1})$  to experience  $e$ 
13:      Save  $e$  to replay memory  $m$ 
14:      Randomly batch_size from  $m$ 
15:       $S_t \leftarrow S_{t+1}$ 
16:       $t = t + 1$ 
17:      Update network  $\theta$  using Eq. (2)
18:      Set  $\theta' = \theta$ 
19:    end for
20:  end for
21: end

```

4. Experiment

In this section, we employ the Simulation of Urban Mobility (SUMO) traffic simulation software to perform simulation with synthetic datasets (2×2 grid-networks), to assess the effectiveness of the proposed cooperative approach for multi-intersection TSC. The simulation for each episode lasts 3600 seconds. Each agent trains for 200 rounds, using a batch size of 32 for training data, a discount factor of 0.99. Additional parameters used for the agent are listed in Table 1.

Table1. Parameters setting of agent

Paramters	Value
Batch-size	32
Memory size	20000
Episode	200
Simulation time	3600s
Optimizer	Adam
Learning rate α	0.0001
Discount factor γ	0.99

The experimental scenario involves 2×2 grid-networks with an area of $350 \times 350 \text{ m}^2$, illustrated in Fig. 4 where each intersection has 8 arriving traffic lanes and 8 departing traffic lanes and a speed limit of 20 meters per second. The spacing between adjacent intersections is determined to be 200 meters. To enhance the simulation's realism, vehicle generation occurs randomly, the intersection area is reached by vehicles that appear randomly from the entrance lanes.



Figure 4. Simulation grid networks in SUMO

In the simulation environment described above, the proposed approach in this paper is compared with the without-cooperation *DDQN* algorithm and fixed algorithm to illustrate the efficacy of the proposed approach. The bar charts in Fig. 5 display the performance details of each agent under each approach in four metrics during an episode. Significantly, the *DRL*-based control approaches (our approach and without-cooperation) demonstrate greater control efficacy compared to the fixed algorithm. This underscores the clear benefits of using *DRL*-based control methods to tackle traffic signal challenges. Clearly all agents of our approach demonstrate the highest performance for different metrics and can acquire learn strategies to enhance traffic flow at intersections.

All results being perceived for *AWT*, the proposed approach demonstrates a 19.3% improvement of fixed algorithm and a 12.2% enhancement over the without-cooperation algorithm. In *API*, Improvements of 23.2% and 15.3% are attained, respectively. Enhancements of 25.12% and 19.1% are showed in *AEC*. In *ANE*, enhancements of 14.9% and 11.5% are respectively achieved. Besides, the simulation results indicate that the proposed approach enhances the smoothness of multi-intersection traffic.

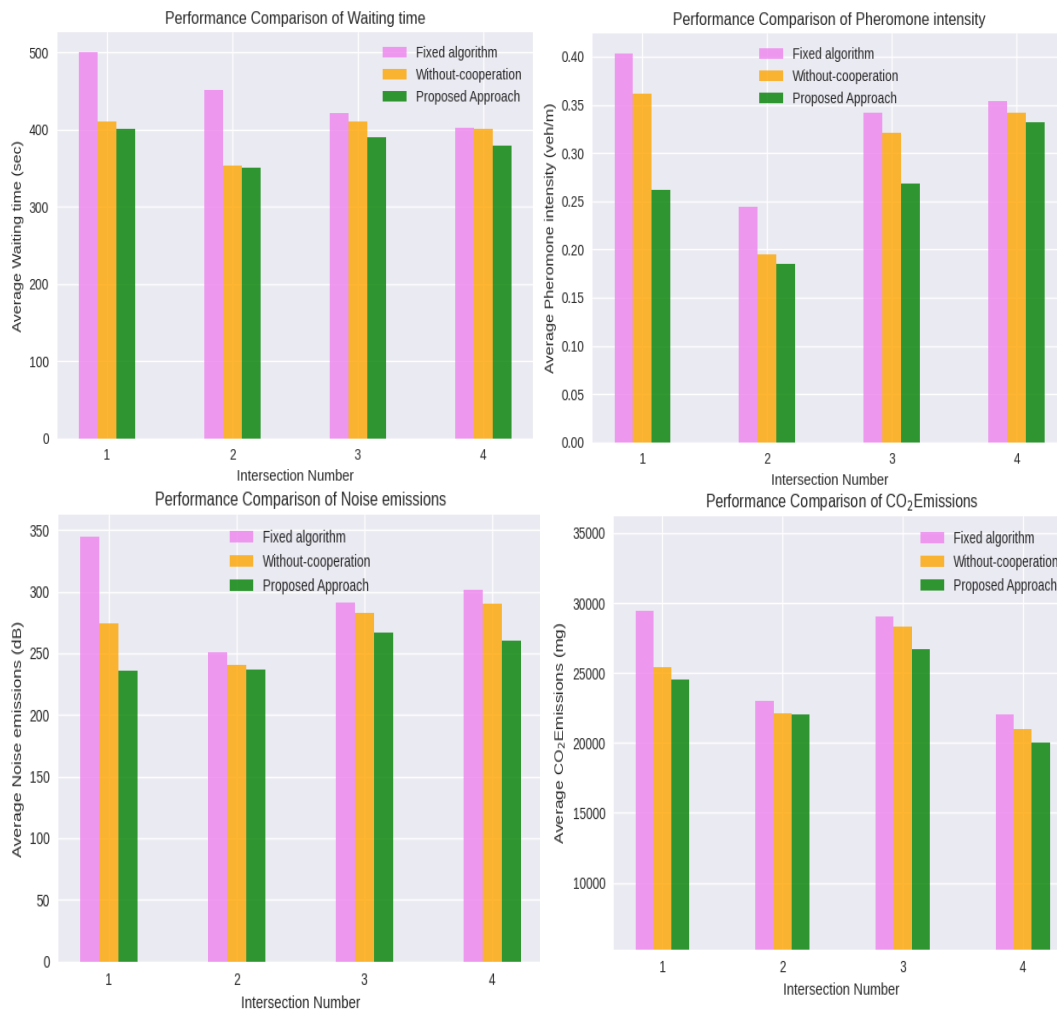


Figure 5. Evaluation results bar charts on each intersection of four metrics

5. Conclusion

In this research, a novel cooperative control scheme based on MARL is introduced to tackle the challenge of multi-intersection traffic control. Our approach takes into account traffic conditions at both the local and neighboring intersections to enhance the visibility of each agent. A novel reward function incorporating global vehicle traffic flow is devised to promote cooperation between

intersections. Also, new road pheromone information-based state is proposed. Simulation results conducted on 2×2 grid networks show that the proposed approach outperforms the fixed and without-cooperation algorithms for different traffic metrics. Furthermore, in future research, we plan to integrate pedestrian information to ensure equitable passage for all intersection users into ATSC algorithm.

References

- [1] T. A. Haddad, D. Hedjazi, and S. Aouag, "An IoT-Based Adaptive Traffic Light Control Algorithm for Isolated Intersection," in *Advances in Computing Systems and Applications*, vol. 199, M. R. Senouci, M. E. Y. Boudaren, F. Sebbak, and M. Mataoui, Eds., in Lecture Notes in Networks and Systems, vol. 199, Cham: Springer International Publishing, 2021, pp. 107–117. doi: 10.1007/978-3-030-69418-0_10.
- [2] C. Cai, C. K. Wong, and B. G. Heydecker, "Adaptive traffic signal control using approximate dynamic programming," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 5, pp. 456–474, Oct. 2009, doi: 10.1016/j.trc.2009.04.005.
- [3] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction.," *MIT press*, 2018.
- [4] T. A. Haddad, D. Hedjazi, and S. Aouag, "A New Deep Reinforcement Learning-Based Adaptive Traffic Light Control Approach for Isolated Intersection," in *2022 5th International Symposium on Informatics and its Applications (ISIA)*, M'sila, Algeria: IEEE, Nov. 2022, pp. 1–6. doi: 10.1109/ISIA55826.2022.9993598.
- [5] W. Genders and S. Razavi, "Using a Deep Reinforcement Learning Agent for Traffic Signal Control," Nov. 03, 2016, *arXiv: arXiv:1611.01142*. Accessed: Jan. 14, 2023. [Online]. Available: <http://arxiv.org/abs/1611.01142>
- [6] J. Gao, Y. Shen, J. Liu, M. Ito, and N. Shiratori, "Adaptive Traffic Signal Control: Deep Reinforcement Learning Algorithm with Experience Replay and Target Network," *arXiv:1705.02755 [cs]*, May 2017, Accessed: Mar. 16, 2020. [Online]. Available: <http://arxiv.org/abs/1705.02755>
- [7] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Multiagent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC): Methodology and Large-Scale Application on Downtown Toronto," *IEEE Trans. Intell. Transport. Syst.*, vol. 14, no. 3, pp. 1140–1150, Sep. 2013, doi: 10.1109/TITS.2013.2255286.
- [8] D. Kim and O. Jeong, "Cooperative Traffic Signal Control with Traffic Flow Prediction in Multi-Intersection," *Sensors*, vol. 20, no. 1, p. 137, Dec. 2019, doi: 10.3390/s20010137.
- [9] T. A. Haddad, D. Hedjazi, and S. Aouag, "A deep reinforcement learning-based cooperative approach for multi-intersection traffic signal control," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105019, Sep. 2022, doi: 10.1016/j.engappai.2022.105019.
- [10] P. La and S. Bhatnagar, "Reinforcement Learning With Function Approximation for Traffic Signal Control," *IEEE Trans. Intell. Transport. Syst.*, vol. 12, no. 2, pp. 412–421, Jun. 2011, doi: 10.1109/TITS.2010.2091408.
- [11] S. S. Mousavi, M. Schukat, and E. Howley, "Traffic light control using deep policy-gradient and value-function-based reinforcement learning," *IET Intelligent Transport Systems*, vol. 11, no. 7, pp. 417–423, Sep. 2017, doi: 10.1049/iet-its.2017.0153.
- [12] K. Zhang, Z. Yang, and T. Başar, "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," in *Handbook of Reinforcement Learning and Control*, vol. 325, K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, Eds., in Studies in Systems, Decision and Control, vol. 325, Cham: Springer International Publishing, 2021, pp. 321–384. doi: 10.1007/978-3-030-60990-0_12.
- [13] T. Chu, J. Wang, L. Codeca, and Z. Li, "Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 3, pp. 1086–1095, Mar. 2020, doi: 10.1109/TITS.2019.2901791.
- [14] T. A. Haddad, "Une approche coopérative basée sur l'IoT pour améliorer la qualité du trafic routier," Thèse de Doctorat, Université de Batna 2, 2023. [Online]. Available: <https://dspace.univ-batna2.dz/handle/123456789/1799>
- [15] D. Srinivasan, M. C. Choy, and R. L. Cheu, "Neural Networks for Real-Time Traffic Signal Control," *IEEE Trans. Intell. Transport. Syst.*, vol. 7, no. 3, pp. 261–272, Sep. 2006, doi: 10.1109/TITS.2006.874716.

- [16] C. Wan and M. Hwang, "Value-based deep reinforcement learning for adaptive isolated intersection signal control," *IET Intelligent Transport Systems*, vol. 12, no. 9, pp. 1005–1010, Nov. 2018, doi: 10.1049/iet-its.2018.5170.
- [17] H. Ge, Y. Song, C. Wu, J. Ren, and G. Tan, "Cooperative Deep Q-Learning With Q-Value Transfer for Multi-Intersection Signal Control," *IEEE Access*, vol. 7, pp. 40797–40809, 2019, doi: 10.1109/ACCESS.2019.2907618.
- [18] C. Chen *et al.*, "Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control," *AAAI*, vol. 34, no. 04, pp. 3414–3421, Apr. 2020, doi: 10.1609/aaai.v34i04.5744.
- [19] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," *In Proceedings of the tenth international conference on machine learning*, p. (pp. 330-337), 1993.
- [20] J. A. Calvo and I. Dusparic, "Heterogeneous Multi-Agent Deep Reinforcement Learning for Traffic Lights Control.," *AICS*, pp. 2–13, 2018.
- [21] S. Yang and B. Yang, "A Meta Multi-agent Reinforcement Learning Algorithm for Multi-intersection Traffic Signal Control," in *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCCom/CyberSciTech)*, AB, Canada: IEEE, Oct. 2021, pp. 18–25. doi: 10.1109/DASC-PiCom-CBDCCom-CyberSciTech52372.2021.00019.
- [22] T. Tan, F. Bao, Y. Deng, A. Jin, Q. Dai, and J. Wang, "Cooperative Deep Reinforcement Learning for Large-Scale Traffic Grid Signal Control," *IEEE Trans. Cybern.*, pp. 1–14, 2019, doi: 10.1109/TCYB.2019.2904742.
- [23] C. J. C. H. Watkins, "Learning from Delayed Rewards," *Ph.D. thesis. University of Cambridge, Cambridge, England*, 1989.

Empowering Customer Service with AI: A Multilingual Chatbot for Mobile Companies

Tahar Mehenni^{1✉}, Samiha Attaoua², Chaima Guesmia³

¹ *Laboratory of Informatics and Its Application of M'sila (LIAM), Computer Science Department,
University of M'sila, Algeria*

^{2,3} *Computer Science Department, University of M'sila, Algeria*

Abstract

Nowadays, people are using more and more advanced technology in communication. Consequently, the brand-customer relationship has never been more intense, thanks to rapid technological advancements. Customers are preferring self-service options satisfying autonomy and enabling them to make purchases or access information without engaging with human representatives. Conversational systems (chatbots) present the most interesting solution to both enhancing customer-brand communication and improving the customer experience. Mobile companies are considering this technology in order to perform a better engagement with clients using software platforms, that can offer regular chat functionalities, i.e. giving more information, solving problems, making appointments, and more other advanced features. Moreover, with the presence of global customer bases, mobile operators need conversational systems that can handle queries using multiple languages. Therefore, implementing chatbots seems to be an effective and profitable solution for both parties. These systems have to be designed to simulate call center employees by responding to common inquiries and guiding customers in their languages, in order to reduce the long time spent searching for information. In this paper, we present a rule-based chatbot for mobile company customers that supports three languages: Arabic, English and French, and automatically recognizes the user's language without any additional information. The developed prototype is evaluated via a pre-defined questionnaire and the results are promising, indicating that this solution could be advantageous for mobile companies and other organizations.

Keywords: Multilingual, Rule-based Chatbot, Mobile Company, Customer service, Natural Language Processing (NLP).

1. Introduction

Artificial intelligence (AI) is becoming an important part of our daily tasks through the development of several intelligent systems and devices, commonly known as intelligent agents. These agents can perform a wide range of tasks, spanning from basic works to complex operations. Chatbots can be defined as intelligent agents that perform interaction between computers and humans, representing a prominent example of intelligent human-computer interaction (HCI).

A chatbot is a software designed to handle conversational interactions through text or voice and simulate intelligent behavior while performing these interactions using Natural Language Processing (NLP) techniques. Therefore, a chatbot is an interactive agent, a digital assistant, or an artificial conversational entity, facilitating interactions without the need for direct human involvement (Figure 1 illustrates the key components of a chatbot system).

✉ Corresponding Author Email: tahar.mehenni@univ-msila.dz

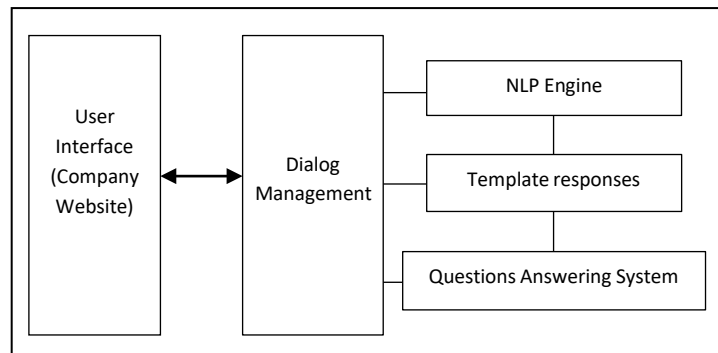


Fig. 1. Main components of a chatbot system.

In the past, customer support was primarily handled through phone calls and later expanded to email communication. This was performed with the presence of human agents providing responses to client requests. However, as the capacity of human support teams is often limited, especially for multiple brands and big companies, the challenge of providing efficient and in-time support persists. In fact, customers demand more and more self-service options due to extended wait times and complex administrative process.

To improve availability and meet customer prospects, companies are transitioning from human-to-human interactions towards self-service models or computer-to-human solutions. They are using chatbots as automated customer support agents available 24/7 on software platforms acting as the best communication channels preferred by customers.

Last decades have known a revival of interest in chatbot technology. This progress can be attributed to increased computational power, the availability of vast datasets, and the advanced of NLP and deep learning techniques. As a result, chatbots have become adopted in many domains and applied in several companies.

The objective of this study is to propose a solution to address the challenges faced by a mobile company and its customers, meeting their expectations. This solution involves a chatbot system that:

- Responds quickly and efficiently to customer inquiries with relevance and confidentiality.
- Enables customers to express complaints in the language they know in order to find solutions or establish resolution guidelines.
- Offers company services with new promotions and offerings.

The remaining of this paper is organized as follows. Section 2 presents related work in chatbot domain. In section 3, a description of the proposed chatbot system is given. Evaluation results and discussion are presented in section 4 and finally, the paper is concluded in section 5.

2. Related Work

Customers of a company are increasingly seeking autonomy and self-service options to obtain information, resolve problems or make purchases without the need of interacting directly with brand representatives and thus, spending time and money. To tackle these requirements, companies are replacing human-to-human interaction by sophisticated computer-to-human systems. Therefore, integrating chatbots into customer service represents an interested solution to handle efficiently customer-brand communication.

Companies are developing this advanced technology to empower engagement with their clients, using software platforms that can offer several features and various advanced capabilities.

Deep learning enables models composed of multiple processing layers to learn representations of data with multiple levels of abstraction. Deep learning finds complex structures within voluminous datasets using the backpropagation technique to adjust internal parameters of the machine. Deep convolutional networks have revolutionized image, video, speech, and audio processing, while recurrent networks have enhanced sequential data understanding, including text and speech [1,2].

Speech or text interactions between humans and computers are more and more common. While people would like to communicate with computers in ways similar to how they interact with humans, NLP comes to play a critical role in analyzing speech and delivering human-like responses.

Chatbot architectures typically are divided into two categories: rule-based systems and corpus-based systems [2]. Rule-based systems contain primary powerful chatbots, while corpus-based systems involve large datasets of human-human conversations, employing techniques such as information retrieval or neural network sequence-to-sequence systems to map user statements to system responses. In [3], researchers developed a neural conversation model (NCM) using a sequence-to-sequence approach. They trained their chatbot uniquely on one dialogue corpus and evaluated its performance against Cleverbot [4], demonstrating higher performance. Another dialogue modeling approach proposed by [5] focuses on answering questions within a specific domain. Their method involves dividing dialogue corpora into expression and response pairs, searching for the most similar expression to a given question within the dataset, and providing the corresponding response. They evaluated their model by allowing users to rate the chatbot's answers based on effectiveness.

Several multilingual chatbots are available on the web, like Mondly that supports 30 languages, Memrise with 30 and Watson with 21 languages. Unfortunately, these chatbots are not free [6]. Authors in [7] developed a program using Microsoft Visual Studio that linked Google Translate with a free chatbot, called Tutor Mike, thus allowing users to chat in any of 103 different languages. However, this system is evaluated by students using only 4 languages.

Many chatbots have been developed for the customer service domain in different companies. Authors in [8] proposed a multilingual chatbot system that allow companies and organizations to customize and deploy their own multilingual chatbot service. The chatbot recognizes English, Malay and Chinese. Additional details about using chatbot technologies within customer service operations can be found in [9].

Our proposed chatbot has been designed and developed to tackle the diverse needs of customers served by an important telecommunications company in Algeria. This conversational system is designed to provide smooth multilingual support, accommodating the linguistic diversity of the region and the global customer base. Using advanced NLP techniques, the proposed chatbot offers a user-friendly interface where customers can easily communicate their inquiries, concerns, and issues.

3. The proposed chatbot system

A rule-based system applies human-made rules to store, sort and manipulate data simulating human intelligence. Therefore, rule-based systems require a set of facts or source of data, and a set of rules for manipulating that data. The rule-based chatbot answers the customer queries based on some predefined rules, therefore, its implementation is very straightforward and cheap [10].

While AI conversational chatbots using corpus based systems and machine learning techniques are more advanced, they're not always necessary. For smaller companies or those with specific goals, rule-based chatbots can give more appropriate solution.

Given the constraints of operating within a specific closed domain related to the mobile company, the most suitable approach was to employ a rule-based chatbot system. The proposed chatbot is implemented for a specific category of users which are the company customers, meaning that most conversations and customer inquiries of such users can be expected.

Additionally, rule-based chatbots offer several other beneficial aspects, which can be outlined as follows:

- In contrast to their AI-powered counterparts that need huge datasets and computation resources for training, rule-based chatbots can be implemented in a short period through defining certain rules and responses. As such, they can be extremely inexpensive for businesses to use, especially small to medium size enterprises with a limited technology budget.
- Since they work on predefined rules and structured data formats, they can be configured and implemented in different environments and enterprises without requiring big updates or modifications or even significant development.
- They can be configured with clear triggers that can find with precision when a conversation should be heightened to a human agent. These triggers can be simply keywords or query patterns. The chatbot can transfer the conversation with the interaction history that the human agent try to pick up exactly.
- The decision-making processes of rule-based chatbots are based on predefined rules rather than complex algorithms. Therefore, it should be easier to audit their responses and identify exactly why certain answers were given. In that way, they are suitable for financial services, where every interaction must be compliant with security protocols.
- They can be implemented integrating images, videos and other multimedia sources at any points along the conversation. This multimedia capability gives a rich content and valuable information to the users.

The development of the proposed rule-based chatbot typically began with a detailed analysis of the intended use case and target audience. This initial phase involved collaborating closely with subject matter experts to identify common user queries, map out conversation flows, and define the chatbot's scope. Once the initial work is achieved, we use Dialogflow as a specialized development platform to create the chatbot's structure.

Dialogflow is a natural language understanding platform used to design and integrate a conversational user interface into mobile apps, web applications, devices, bots, interactive voice response systems and related uses. This development tool is a part of Google Cloud Platform since 2017 [11]. It provides intuitive interfaces for designing conversation trees, defining intents, and creating entities. The core of a rule-based chatbot lies in a set of if-then rules and pattern-matching algorithms, which determine how the chatbot responds to specific user inputs.

The training process for a rule-based chatbot differs significantly from that of machine learning-based chatbots. Instead of learning from vast datasets, rule-based chatbots rely on manually defined rules and patterns. We build these rules from the telecom domain, often using regular expressions to capture various ways users might phrase their queries. This process involved extensive testing and iteration, where we simulate conversations to identify gaps in the chatbot's knowledge base and refine its responses.

Natural Language Processing (NLP) techniques were used to provide chatbots the ability to understand user inputs, even when faced with variations in phrasing or spelling errors expressed with multiple languages. We used NLTK Tool (Natural Language Toolkit) to implement these NLP capabilities [12]. NLTK helps in providing many advanced features to perform easily natural language processing:

1. Lemmatization and Stemming: These features can transform words to their base or root form, in order to easily analyze texts.
2. Grammar Parsing: NLTK helps in building parse trees to understand the grammatical structure of sentences.
3. Part-of-speech (POS) using NLTK helps in assigning grammatical parts of speech to each word in a text, such as noun, verb, adjective, or adverb. This is important to better understanding sentences and the relationships between words.
4. Named Entity Recognition (NER) in NLTK is an important feature that identifies and classifies key information in text into predefined categories such as names of persons, organizations, locations, expressions of times, quantities, monetary values, and percentages.

The implementation of these NLP features using NLTK tool can provide chatbot systems with many strong capabilities in order to better understand and respond to user complaints and requests making interactions more natural.

We have relied in building our rule-based chatbot on the following steps [13,14]:

- (1) Import the packages/libraries.
- (2) Create a function that calculates the probability of the corresponding message.
- (3) Create a variable that determines certainty message.
- (4) Check that the required words are also included.
- (5) Create a return statement for the function that returns the precision of each sentence.
- (6) Generate a key in the bot response index.
- (7) Create our response so that the key is the response.

As a result, we created a chatbot that automatically recognizes the customer's language and gives answers to it automatically in the same language so that it responds to common questions for the company customers.

With our rule-based chatbot, one can write inquiries with any language he/she wants, in the same session, and the chatbot recognizes the language and answers with the same language. The languages offered are: Arabic, French and English. However, the chatbot can be expanded to recognize any other language. Fig. 2 shows a screenshot for the developed chatbot.

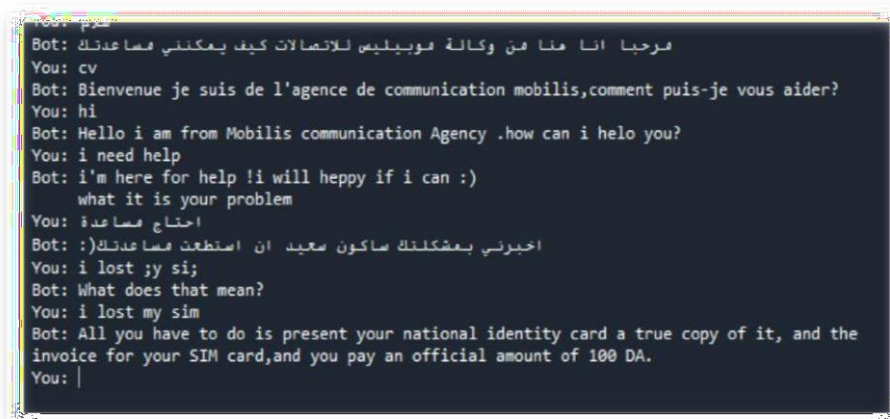


Fig. 2. A screenshot for the proposed chatbot of a mobile company

4. Experimentation results and discussion

We implemented the chatbot and integrated it in the site of the company. The input-output features are shaped in order to improve the efficiency of the system. This application acts as human replacement for proper guidance or information retrieval.

In the beginning, we have created targets for each type of frequently asked questions (various offers and inquiries) that the company customers ask. Then we gave them the possible expressions grouped correctly within well-defined goals and the appropriate answers to them. We evaluated the rule-based chatbot using a questionnaire [15]. A total of 113 people participated to the questionnaire along 5 weeks from a total of 387 invitations sent out in the Telecom Company.

The questionnaire is composed of 7 questions addressed to the customers using the site of the company (see Table 1). Question 1 is related to the language used and clearness of sentences. Question 2 emphasizes the effectiveness of answers given by the chatbot. Questions 3 to 6 are related to the quality of the chatbot responses, and finally Question 7 gives a rate for the received customer support. Answers of all the questions are rated using 1-5 scale.

N°	Question
Q1	How easy or difficult was it to understand the answers of our live chat?
Q2	Did our live chat address your concerns clearly and effectively?
Q3	How satisfied are you with the promptness of our live chat in addressing your concerns?
Q4	How satisfied are you with the resolution provided for your queries or issues?
Q5	How would you rate the courtesy and professionalism of our live chat?
Q6	How would you rate the knowledge and expertise of our live chat?
Q7	How satisfied are you with the customer support you received from our live chat?

Table 1. The customer survey questionnaire for the mobile company.

As shown in Figures 3 and 4, our chatbot seems giving understandable and effective answers in the same language of the questions given by the customers. In fact 50% of the respondents found the answers easy or very easy to understand and agree their effectiveness.

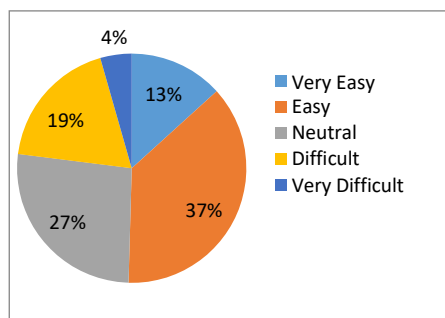


Fig. 3. Understanding the answers of the live chat

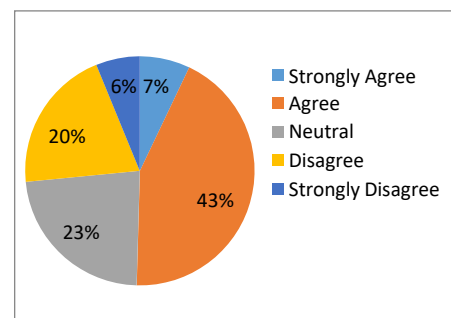


Fig. 4. Effectiveness of the live chat.

The Figures 5 to 8 emphasize the quality of the chatbot w.r.t. the given responses to the customers. Four ratings are selected: promptness, resolution, professionalism and expertise. As shown in these figures, the respondents give high rates (52% to 77%) to the chatbot they used. Especially, the chatbot demonstrated its high promptness and expertise in replying the customers. Finally, according to the Fig. 9, most respondents (66%) are satisfied with the customer support they received using the chatbot.

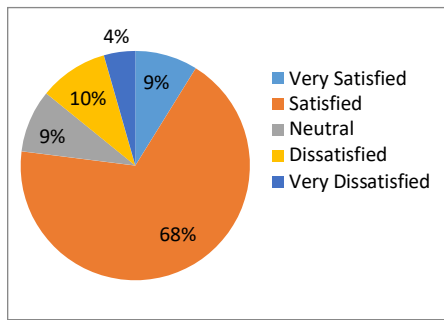


Fig. 5. Promptness of the live chat in addressing customer concerns

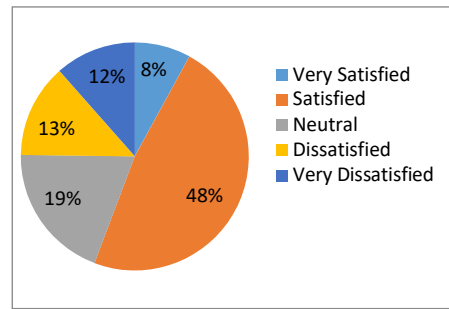


Fig. 6. Resolution provided by the live chat

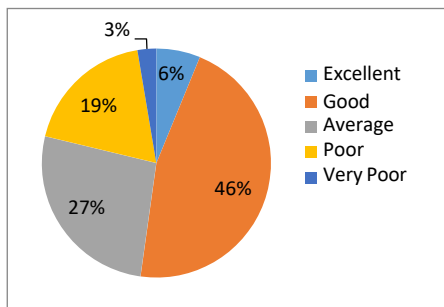


Fig. 7. Courtesy and professionalism of the live chat.

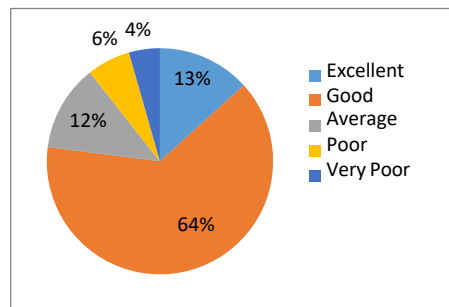


Fig. 8. Knowledge and expertise of the live chat.

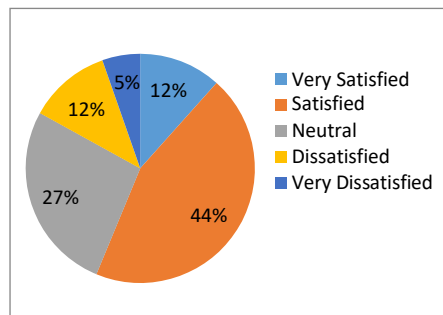


Fig. 9. Overall rate of the customer support received from the live chat.

Conclusion

In this paper, we proposed a multilingual chatbot in order to answer common questions for telecom customers without having to contact with a direct agent. The chatbot is available all the time to provide assistance, and it is easy to integrate into the help pages on the company's websites.

We aimed to create a chatbot based on artificial intelligence and natural language processing. Since we work in a closed field and the bot we want in the end is dedicated to answering common questions, we proposed a Rule-based chatbot provided with information knowledge according to the user's need. The developed rule-based chatbot was able to recognize three languages, which are English, French and Arabic in the same session.

The proposed chatbot is equipped to handle a wide array of customer problems with precision and efficiency, i.e. addressing technical difficulties, billing questions, service upgrades, or general information requests. Its ability to understand and respond in multiple languages ensures that language barriers are eliminated, making customer support accessible to all segments of the Algerian population and the tourists. This innovative solution not only enhances the overall customer experience but also updates the company's support operations, reducing wait times and improving resolution rates for customer inquiries.

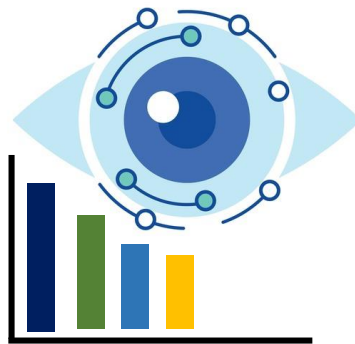
For future work, we will attempt to create a chatbot using new and different methods of deep learning. So that it achieves the desired result by covering most of the customers' interests with telecom companies. Generative models (Transformers based models) are well known for their capability to overcome rule based issue like deep context understanding.

References

1. Krizhevsky, A., Sutskever, I., G. Hinton, E., (2012) ImageNet Classification with Deep Convolutional Neural Networks. In "Proceedings of the 25th NIPS Conference".
2. Lee, L.J., Frey, B.J., Leung, M.K., Xiong, H.Y., (2014) Deep learning of the tissue regulated splicing code, *Journal of Bioinformatics*.
3. Vinyals, O., Quoc, V., (2015) A neural conversational model. CoRR, abs/1506.05869.
4. Carpenter, R., A Clever Bot, <http://www.cleverbot.com>, (last accessed Mars 13, 2024).
5. Ameixa, D., Coheur, L., Fialho, P., Quaresma P., Luke, (2014) I am your father, 14th International Conference, Boston, MA, USA.
6. Mohammad, Z., (2018) Build multilingual chatbots with Watson Language Translator & Watson Assistant. <https://medium.com/ibmwatson/>.
7. Vanjani, M., Aiken, M., & Park, M., (2019) Chatbots for multilingual conversations. *Journal of Management Science and Business Intelligence*, 4(1), pp. 19-24
8. Kasinathan, V., Mustapha, A., & Bin, C. K., (2021) A Customizable multilingual chatbot system for customer support. *Annals of Emerging Technologies in Computing (AETiC)*, 5(5), pp. 51-59.
9. Nirala, K. K., Singh, N. K., & Purani, V. S., (2022) A survey on providing customer and public administration based services using AI: chatbot. *Multimedia Tools and Applications*, 81(16), pp. 22215-22246
10. Thorat, S. A., & Jadhav, V. (2020). A review on implementation issues of rule-based chatbot systems. In *Proceedings of the international conference on innovative computing & communications (ICICC)*.
11. Muhammad, A. F., Susanto, D., Alimudin, A., Adila, F., Assidiqi, M. H., & Nabhan, S. (2020). Developing English conversation chatbot using dialogflow. In *2020 International Electronics Symposium (IES)* pp. 468-475. IEEE.
12. NLTK: Natural Language Toolkit, <https://www.nltk.org>. (last accessed Mars 26, 2024).
13. Habibur, R., Anytime Customer Support using Artificial Intelligence Chatbot in Customer Service, <https://www.codeproject.com/Articles/5274317/Anytime-Customer-Support-using-Artificial-Intellig>, (last accessed Mars 18, 2024).
14. Usman, S., Building a Rule-based Chatbot in Python, <https://online.datasciencedojo.com/blogs/rule-based-chatbot-in-python>, (last accessed Mars 18, 2024).
15. Maroengsit W., Piyakulpinyo T., Phonyiam K., Pongnumkul S., Chaovalit P. and Theeramunkong T. (2019) A Survey on Evaluation Methods for Chatbots. In *Proceedings of the 7th International Conference on Information and Education Technology*, pp. 111-119

SESSION 5

Data Science, Computer Vision, and Image Processing



Pi-Based Scrambling Algorithm for Secure Image Watermarking

Nesrine Benmeheni^{1✉}, Abdelhamid Benhocine¹, Nasser Eddine Mouhoub²

¹ Ferhat Abbas University, Setif1, Mathematical Departement, LMFN Laboratory, Setif, Algeria
{ nesrine.benmehenni, abenhocine }@univ-setif.dz

² M'sila University, College of Mathematics and Computer Science, M'sila, Algeria
nassereddine.mouhoub@univ-msila.dz

Abstract

In this work, we present an image scrambling algorithm as part of a medical image watermarking scheme. The proposed scrambling technique is inspired by the well-known Arnold transform, which is commonly used in image processing for its chaotic properties. However, our approach introduces a novel variation by using sequences of digits derived from the number $\mathbf{\Pi}$ to control the scrambling process. The use of $\mathbf{\Pi}$ a non-repeating irrational number, ensures that the scrambling is both highly unpredictable and deterministic, providing a robust method for securing sensitive medical images. This scrambling mechanism enhances the security of the watermarking process by adding an additional layer of obfuscation, making the embedded watermark more resistant to tampering and unauthorized extraction.

The integration of this scrambling algorithm into the watermarking framework provides a significant improvement in the security of medical image protection, which is critical for maintaining patient confidentiality and data integrity in healthcare systems.

Keywords: Scrambling, $\mathbf{\Pi}$, Wtermarking

1. Introduction

Nowadays, with the digitization and centralization of information systems and access to patients' medical data in hospitals, as well as the increase in digital exchanges in contexts such as telemedicine, clinical research, and computer-assisted diagnosis, data protection has become crucial and necessary. One of the proven techniques for ensuring this protection is watermarking[1][2][3], which helps safeguard the integrity, authenticity, and confidentiality of medical images. Watermarking allows information to be invisibly embedded in medical images, such as patient identifiers, metadata, or copyrights, without affecting their quality or clinical utility[4][5][6][7][8]. This information can take the form of text or an image. However, to prevent any attempt at unauthorized disclosure or extraction of the hidden information, it is essential to scramble these marks (scramble the embedded image in our work). Scrambling makes the watermark more resistant to attacks by making it even more difficult to extract. Among the most well-known scrambling algorithms is Arnold scrambling, which is based on chaotic pixel transformation and is particularly effective. This process disorganizes the spatial structure of the image through precise iterations, rendering the hidden information inaccessible without the correct decoding parameters[9][10][11][12][13][14][15]. Our work consists of adding a scrambling algorithm inspired by Arnold's scrambling, using sequences of digits derived from the number pi, to our

watermarking scheme to significantly enhance the security of medical images. We name this algorithm Pi-Key, which is explained in the following sections.

2. Watermarking for medical image

Watermarking is essential in the medical field as it helps protect sensitive patient data and ensures the integrity of medical images. With the rise of digital imaging and the centralization of hospital data, the risk of hacking increases, potentially allowing unauthorized access to personal or compromising information. For instance, in Algeria, medical images, such as MRI scans, are often sent to specialists abroad for diagnosis, highlighting the need for secure transmission of these sensitive files.

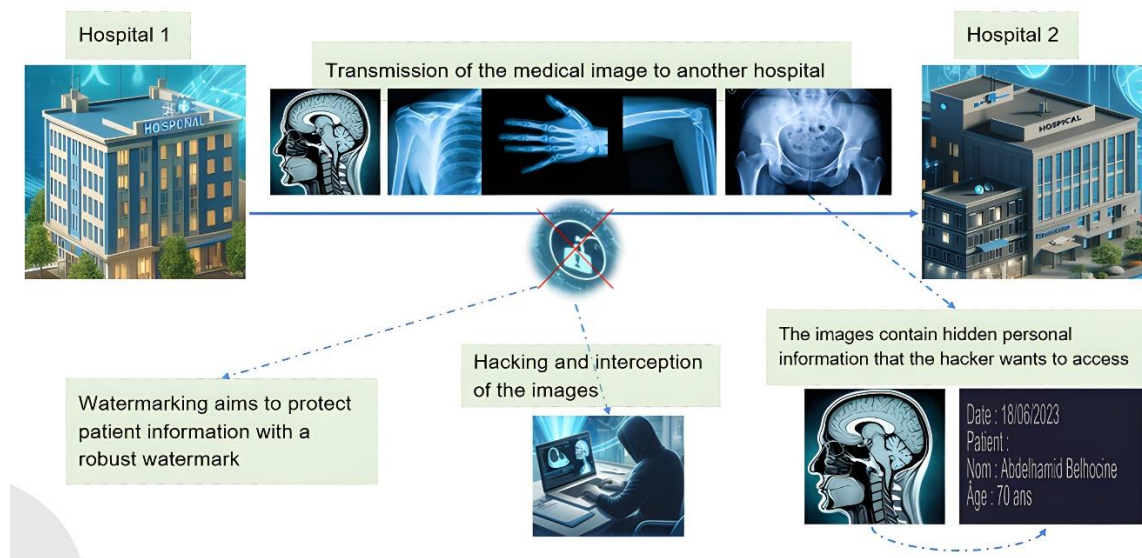


Figure 1. Importance of watermarking in medical field

2.1 General watermarking scheme with scrambling the watermark image

Now, we're focusing on making personal data in watermarks more secure. Our plan involves incorporating a method for encrypting images, in our recent watermarking schemes. This method works by rearranging pairs of pixels to create a new image that's the same size. In this study, we will encrypt the watermark image using a scrambling algorithm, to encrypt our watermark such as:

Arnold Cat Map, Arnold Scrambling, Chaos-Based Scrambling, Block Permutation, Pseudo-Random Scrambling, Bit-Level Scrambling...

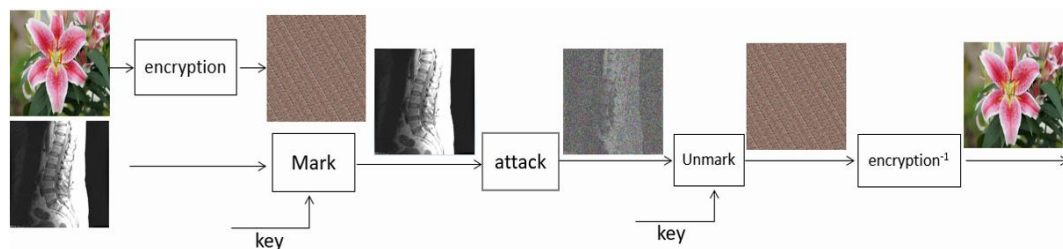


Figure 2. General watermarking scheme using an encrypted imbedding image

3. Scrambling an image :

Image scrambling refers to the rearrangement of an image's pixels according to a defined procedure. Generally, scrambling is reversible, allowing one to return to the original image after scrambling. Among the most well-known methods are Arnold's transform, the Fibonacci transform, 2D mapping, key-based row and column shifting, gray code with bit plane transformation, and key-based row and column shifting combined with bit-level permutation[9].

3.1 Swapping blocs of an image:

In this section, we present a simple example of scrambling, which involves dividing an image into four parts and swapping the block in the top left with the block in the bottom right, and the block in the top right with the block in the bottom left. This operation can be performed with a number of blocks equal to 4, 16, 64, etc. See Figure 3 for the result.

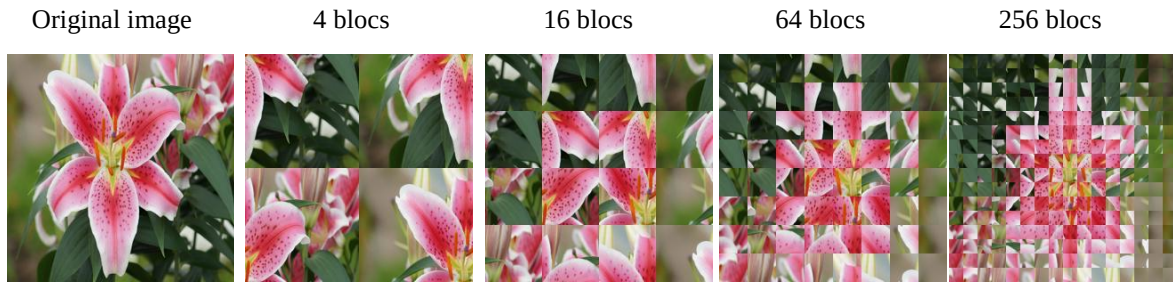


Figure 3. Swapping blocs of an image depending of the number of blocs

3.2 Arnold scrambling:

Arnold scrambling is an image transformation method that rearranges pixel positions using a series of permutations and rotations. It operates according to a specific mathematical formula, providing reversible scrambling of images for security purposes.

The Arnold scrambling transformation formula for a 2D image can be expressed as follows:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} * \begin{pmatrix} x \\ y \end{pmatrix} \bmod N$$

Where (x', y') , are the coordinates of the transformed pixel, (x, y) are the coordinates of the original pixel, and N is the size of the image,

The value of N is determined by the size of the larger dimension of the image, whether it be the height or the width. This ensures that all transformation operations are performed consistently across the entire image.

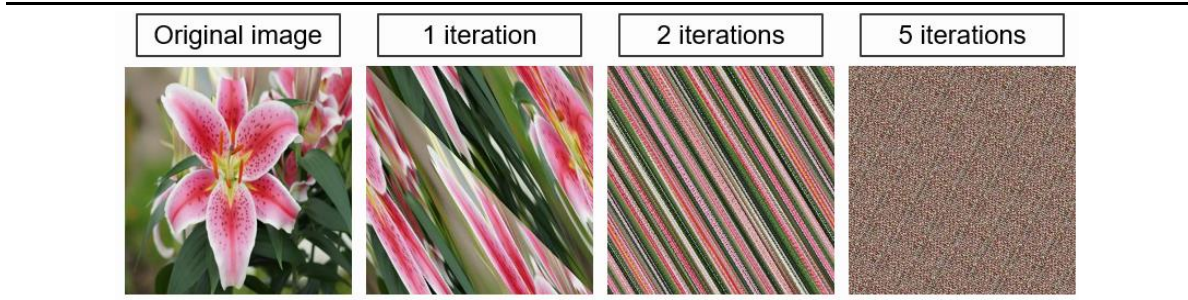


Figure 4. Scrambling an image using Arnold's scrambling

3.3 Proposed scrambling algorithm (Pi-key):

As with Arnold's scrambling, we calculate the new pixel indices (i' , j') from the original indices (i , j), using Pi as a chaotic sequence. The decimals of Pi do not exhibit a repetitive pattern, making it an intriguing choice as a key for encryption and decryption. The formula is as follows:

$$x' = (x * \pi_1) \bmod n \quad (1)$$

$$y' = (y * \pi_2) \bmod m \quad (2)$$

where n is the height of the image, and m is its width. π_1 and π_2 are random numbers taken from Pi, with a specific length. To ensure we can recover the original indices (i , j), π_1 and n must be coprime, as well as π_2 and m .

To compute the modular inverse of $\pi_1 \bmod n$ and $\pi_2 \bmod m$, we use Python's `pow(base, exp, mod)` function.

When $exp = -1$, this function calculates the modular inverse of the base. Hence, `pow(pi_1, -1, height)` computes the modular inverse of π_1 modulo the image height.

$$\pi_1^{-1} = \text{pow}(\pi_1, -1, n) \quad (3)$$

$$\pi_2^{-1} = \text{pow}(\pi_2, -1, m) \quad (4)$$

Compute the original coordinates:

$$x = (x' * \pi_1^{-1}) \bmod n \quad (5)$$

$$y = (y' * \pi_2^{-1}) \bmod m \quad (6)$$

3.3.1 Scrambling an image using Pi-key algorithm:

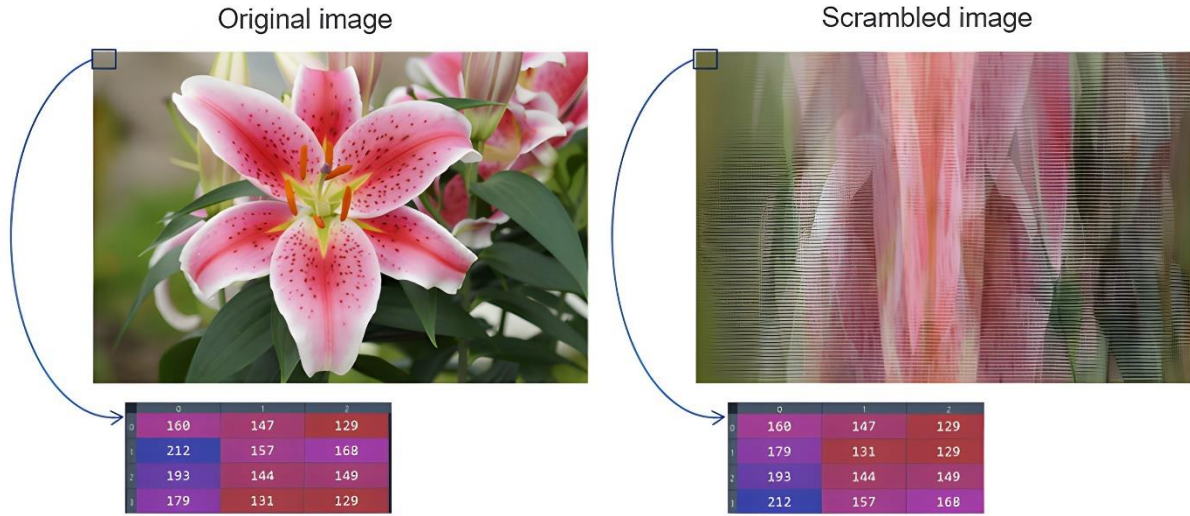


Figure 5. Scrambling an image using Pi-key

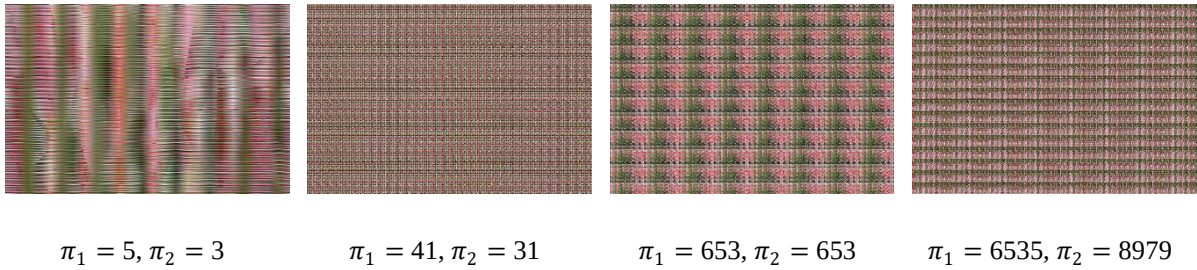


Figure 6. Scrambling an image with different value of π_1, π_2

3.3.2 Recovering the image:

Using Python we generate 2 random integers to determine two positions in our π sequence. These positions are used to identify the values of π_1 and π_2 , where the first random number corresponds to the position for π_1 , and the second random number corresponds to the position for π_2 . Both random numbers are drawn from a uniform distribution, ensuring that every position in π has an equal probability of being selected.

Next, we test whether π_1 and π_2 are coprime with N and M, respectively. If they are not, the random number generation is repeated. For simplicity, let us consider the case where the cardinal π_1 and π_2 is equal to one. These values, along with the current pixel positions (i, j), are used to compute the new pixel location.

Subsequently, using Equations (3) and (4), we calculate the modular inverses π_1^{-1} and π_2^{-1} to revert the pixel positions to their original locations, as shown in Figures 5 and 6.

In this case, we had the values $\pi_1 = 5, \pi_2 = 3$ for scrambling the image, and for recovering the original image from the scrambled one $\pi_1^{-1} = 205, \pi_2^{-1} = 171$. Figure 7 shows that the restored image is identical to the original.



Figure 7. Recovering the Scrambled image

3.3.4 Testing the Pi-key scrambling algorithm:

Taking into account $\pi_1 = 6535$ and $\pi_2 = 8979$, we test our scrambling algorithm. Various cuts are applied, and Gaussian noise with a mean of 0 and a standard deviation of 25 is added. The results are presented below (**Figure 8**). The image recovery is quite successful in these tests. Despite removing an entire quarter of the image, it remains largely recognizable.

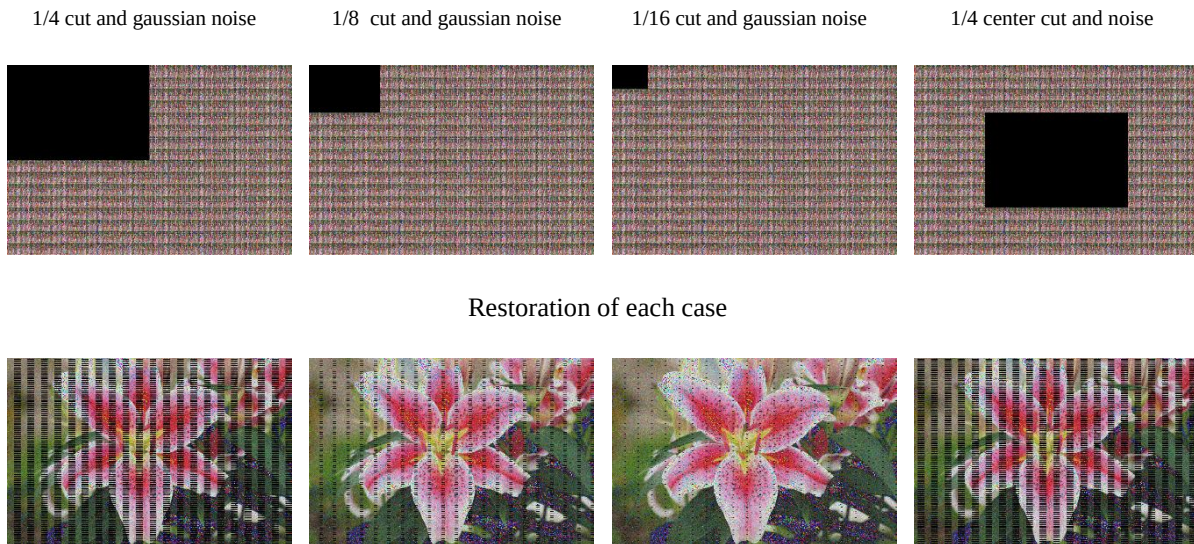


Figure 8. Testing the Scrambling using gaussian noise and various cuts

4. Conclusion

In this work, we presented a scrambling algorithm based on the number pi. The key consists of random numbers that correspond to specific decimals of pi, and these numbers can be combined using multiple randomly generated values to increase the key's complexity.

We tested our scrambling algorithm by subjecting it to attacks involving cuts and Gaussian noise (0, 25). The recovery results are promising, as the image remains recognizable despite the applied transformations. We are continuing to refine and improve the algorithm.

The newly proposed scrambling method yields the desired outcomes. The use of two keys enhances security and provides flexibility in adjusting the size according to specific needs. Our ongoing work aims to further enhance this method and explore its potential applications in cryptography.

Acknowledgements

Thank you to everyone who contributed, directly or indirectly, to this humble work.

References

- [1] A. Anand, J. Bedi, and I. Rida, "MIWET: Medical image watermarking using encryption and fusion technique," *Computers and Electrical Engineering*, vol. 115, p. 109114, 2024.
- [2] O. P. Singh, K. N. Singh, A. K. Singh, A. K. Agrawal, and H. Zhou, "An improved robust algorithm for optimisation-based colour medical image watermarking," *Computers and Electrical Engineering*, vol. 117, p. 109278, 2024.
- [3] A. Tiwari and V. K. Srivastava, "Image watermarking techniques based on Schur decomposition and various image invariant moments: a review," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 16447-16483, 2024.
- [4] A. Benyoucef and M. H. Hamaouche, "Region-based medical image watermarking approach for secure EPR transmission applied to e-Health," *Arabian Journal for Science and Engineering*, vol. 49, no. 3, pp. 4025-4037, 2024.
- [5] U. A. Bhatti, Z. Yu, J. Li, S. A. Nawaz, A. Mehmood, K. Zhang, and L. Yuan, "Hybrid watermarking algorithm using Clifford algebra with Arnold scrambling and chaotic encryption," *IEEE Access*, vol. 8, pp. 76386-76398, 2020.
- [6] S. Gull and S. A. Parah, "Advances in medical image watermarking: a state of the art review," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 1407-1447, 2024.
- [7] R. Singh, R. Pal, H. Mittal, and D. Joshi, "Multi-objective optimization-based medical image watermarking scheme for securing patient records," *Computers and Electrical Engineering*, vol. 118, p. 109303, 2024.
- [8] O. P. Singh, K. N. Singh, A. K. Singh, A. K. Agrawal, and H. Zhou, "An improved robust algorithm for optimisation-based colour medical image watermarking," *Computers and Electrical Engineering*, vol. 117, p. 109278, 2024.
- [9] B. Mondal, "Cryptographic image scrambling techniques," in *Cryptographic and Information Security Approaches for Images and Videos*, J. A. Anderson, Ed. Newbury Park, CA: CRC Press, pp. 37-65, 2018.
- [10] S. Roy and A. K. Pal, "A robust blind hybrid image watermarking scheme in RDWT-DCT domain using Arnold scrambling," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 3577-3616, 2017.
- [11] R. Rhine and N. T. Bhuvan, "Image Scrambling Methods for Image Hiding: A Survey," *International Journal of Computer Science and Network Security*, vol. 15, pp. 86-91, 2015.
- [12] X. Li, "A new measure of image scrambling degree based on grey level difference and information entropy," in *2008 International Conference on Computational Intelligence and Security*, vol. 1, pp. 350-354, Dec. 2008.
- [13] Z. Chen, X. Hou, X. Qian, and C. Gong, "Efficient and robust image coding and transmission based on scrambled block compressive sensing," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1610-1621, 2017.

[14] Z. Chen, L. Li, H. Peng, Y. Liu, H. Zhu, and Y. Yang, "An evaluation method of image scrambling degree based on pixel distribution," in 2018 International Symposium on Security and Privacy in Social Networks and Big Data (SocialSec), pp. 206-211, Dec. 2018.

[15] H. Y. Zhang, "A new image scrambling algorithm," in 2008 International Conference on Machine Learning and Cybernetics, vol. 2, pp. 1088-1092, Jul. 2008.

Parallel Association Rules Mining Using GPUs and Reptile Search Algorithm

Abderrahim Boukhalat^{1,2✉}, KamelEddine Heraguemi³, Mohamed Benouis¹, Brahim Bouderah¹ and Samir Akhrouf¹

¹University of M'Sila, Computer Science Department, M'Sila, Algeria

{Abderrahim.Boukhalat, Mohamed Benouis, Brahim Bouderah, Samir Akhrouf} @univ-msila.dz

² LIM Laboratory, University of Souk Ahras, Computer Science Department, Algeria

³ National School of Artificial Intelligence, Algiers, Algeria

kameleddine.heraquemi@ensia.edu.dz

Abstract

This paper proposes a novel approach to accelerate association rule mining using the Reptile Search Algorithm (RSA) in conjunction with GPU-based parallel processing. Traditional association rule mining techniques can be computationally expensive, especially with large datasets. By utilizing the inherent parallelism of Graphics Processing Units (GPUs), we significantly speed up the fitness evaluation process, a core component of the Reptile Search Algorithm. Our results show a marked improvement in the performance of RSA on large datasets, making it feasible for real-time or large-scale applications such as market basket analysis, healthcare for drug interaction analysis, and web usage mining. We also analyze the impact of various GPU optimizations and present a comparison with CPU-based execution.

Keywords: Data mining, Association rule mining, parallel computing, GPU, Reptile Search Algorithm

1 Introduction

Association Rule Mining (ARM) consists of extracting useful correlations among items in large transactional databases. It is one of the most important tasks in supervised learning. It is used in many fields, including market basket analysis, healthcare for drug interaction analysis, web usage mining, and more. The two most widely used algorithms for ARM are Apriori. [1] and FP-Growth [2]. These are exact algorithms that generate frequent itemsets from which association rules are built. However, generating frequent item sets is an NP-hard problem, making it time-consuming, especially as the size of the dataset grows. Consequently, many metaheuristic algorithms have been proposed in the literature to extract association rules more efficiently. Examples include bee swarm optimization for ARM (BSO) [3], particle swarm optimization (PSO) [4], ant colony optimization (ACO) [5], and Reptile search algorithm [6].

The Reptile Search Algorithm (RSA) is a bio-inspired optimization technique that simulates the hunting and adaptive behavior of reptiles. RSA has been successfully applied to a variety of NP-complete problems across fields such as engineering, finance, and logistics due to its ability to balance exploration and exploitation in search spaces. RSA has been explored in the context of ARM by Abderrahim et al.[7]. Applying RSA-ARM (Reptile Search Algorithm for Association Rules Mining) offers a promising new direction for optimizing rule-mining processes. However, the computational complexity involved in using RSA-ARM on large datasets necessitates innovative approaches.

Association rule mining, especially when combined with metaheuristic algorithms like RSA, is computationally expensive due to the large number of candidate rules that must be evaluated in each iteration of the algorithm. Traditional CPU implementations need help to scale effectively with the increasing size of datasets, especially for ARM tasks combined with complex optimization algorithms like RSA-ARM.[8]. This is due to the inherently sequential nature of CPU computation, which limits parallelism and slows down the process as the dataset grows.[9]. On the other hand, Graphics Processing Units (GPUs)[10], designed for high-throughput parallel computation, provide a promising solution to accelerate the fitness evaluation process by evaluating multiple rules in parallel. Graphics Processing Units (GPUs) offer a powerful solution due to their ability to handle massive parallelism, allowing multiple calculations to occur simultaneously. This makes GPUs well-suited for the high-dimensional computations required by ARM, offering a scalable solution that significantly reduces execution time compared to CPUs.

GPUs have been widely adopted in various computationally intensive tasks, such as deep learning and high-performance computing, due to their ability to perform parallel operations across thousands of cores. Recent studies have demonstrated the significant speedups that can be achieved by parallelizing frequent pattern mining and ARM algorithms on GPUs [11]. To address the computational challenges posed by large datasets, we propose a novel approach that integrates the Reptile Search Algorithm for ARM (RSA-ARM) with GPU parallelism to accelerate computations. In our work, we focus on parallelizing the fitness evaluation function of RSA-ARM, which is the most computationally expensive part of the algorithm. By offloading the fitness evaluation to the GPU using Nvidia CUDA(Compute Unified Device Architecture)[12], we can evaluate multiple candidate rules simultaneously, significantly reducing the time required to mine association rules.

This paper presents the following contributions:

- We present the first implementation of the Reptile Search Algorithm (RSA) for Association Rule Mining (ARM) in GPUs.
- By leveraging the parallel architecture of GPUs, we significantly speed up the fitness evaluation process, enabling RSA to scale to larger datasets.
- We evaluate the performance of the proposed approach on both GPU and CPU, highlighting the speedup achieved through GPU parallelism and analyzing the scalability of the algorithm across different datasets.

The structure of this paper is as follows: Section 2 provides a review of related work on association rule mining, RSA, and parallel GPU computing. Section 3 outlines the methodology, including the integration of RSA-ARM with GPU-based parallelism. Section 4 proposes the parallel approach employed in this study. Section 5 presents the performance evaluation of experimental results and compares the efficiencies of CPU and GPU implementations. Finally, Section 6 concludes the paper and discusses potential directions for future research.

2 Related Work

2.1 Sequential ARM Algorithms:

Association Rule Mining (ARM) has long been studied using sequential algorithms, most notably exact approaches such as Apriori, FP-Growth, DIC, and DHP. These algorithms are highly effective for small to medium-sized datasets but are limited when dealing with large datasets due to their time and space complexity. Apriori, for example, involves multiple passes over the dataset to generate candidate itemsets, making it prohibitively time-consuming for larger datasets. Similarly, FP-Growth, which

compresses the dataset into a compact tree structure, suffers from significant memory overhead when handling large-scale data.

To address these limitations, researchers have explored metaheuristic-based approaches for ARM. Algorithms like Genetic Algorithms for Association Rules (GAR)[13], Particle Swarm Optimization for ARM (PSO-ARM)[14], and Bees Swarm Optimization for ARM (BSO-ARM)[15] Have demonstrated improved performance for large datasets, balancing solution quality and execution time. These methods aim to navigate the vast solution space more efficiently but encounter difficulties with extremely large datasets, such as the WebDocs benchmark, which contains over 1.5 million transactions.

2.2 Algorithms For Parallel Processing:

With the increasing size of datasets, parallelization has become a critical solution for enhancing the performance of ARM algorithms. Early attempts at parallelizing ARM focused on exact methods like Apriori. [16]Agrawal and Shafer introduced parallel Apriori algorithms using strategies like count distribution (CD) and data distribution (DD), where the workload is divided among several processors. Extensions of these methods, such as the Intelligent Data Distribution (IDD) and Hybrid Distribution (HD) models, sought to optimize communication between processors.[17]. However, the limitations of the underlying hardware still bound these approaches.

FP growth has also been parallelized to improve performance on large datasets.[18]. Algorithms such as Multiple Local Frequent Pattern Trees (MLFPT) and distributed FP-Growth achieved some success by distributing the workload across multiple processors and merging local results.[19]. However, the frequent itemset mining task, even when parallelized, remained challenging for large datasets due to the memory requirements of storing FP trees and the computational complexity of pattern discovery.

In addition to exact approaches, parallel metaheuristic algorithms have been developed. Melab et al.[20] They introduced a parallel genetic algorithm for ARM called PGARM, using a master-slave architecture. However, this approach required replicating the entire database on each processor, leading to significant memory usage and potential inefficiency when dealing with very large datasets. Similarly, Mishra et al. [21] IPMOGA divided the search space into regions and assigned them to different processors. However, partitioning the search space efficiently remains a challenge.

In parallel and distributed association rule mining in life science: A novel parallel algorithm to mine genomics data. This paper introduces BPARES (Balanced Parallel Association Rule Extractor from SNPs)[22], a novel algorithm designed to improve the performance of Association Rule Mining (ARM) on biological datasets, particularly those containing Single Nucleotide Polymorphisms (SNPs). Traditional ARM algorithms face significant challenges with execution time, even on small datasets. BPARES leverages parallel computing and a new balancing strategy to optimize response time, efficiently utilizing computational power and reducing memory consumption while maintaining accuracy. The paper also includes a comparison of various sequential, parallel, and distributed ARM techniques.

2.3 GPU resources for parallel processing

With the rise of GPU computing, ARM algorithms have been significantly accelerated by leveraging the massive parallelism that GPUs offer. Wenbin et al. introduced two parallel versions of Apriori for GPU architectures: Pure Bitmap Implementation (PBI) and Trie Bitmap Implementation (TBI) [23] . These implementations transformed datasets into bitmap structures to optimize memory usage and parallelism, achieving speedups of 10x to 15x in their experiments. Syed et al. further improved GPU-based Apriori by implementing a two-step process that first computes itemset support on the GPU before transferring the results back to the CPU for rule generation. This approach showed a 20x speedup but was limited by the cost of CPU-GPU communication. Other GPU-based implementations, such as the

GPAPriori algorithm [24], used compact data structures to accelerate item set counting and reduce memory usage, achieving up to 100x speedups on small datasets. However, as the number of transactions increases, the performance benefits diminish due to thread divergence and memory bottlenecks.

Parallel metaheuristic algorithms have also benefited from GPU acceleration. Cano et al.[25] They proposed a GPU-accelerated evolutionary algorithm for ARM, parallelizing the fitness evaluation step to allow multiple rules to be evaluated simultaneously on the GPU. This approach reduced computation time, but the number of rules that could be evaluated in parallel was limited by the GPU's memory capacity. Strategies to overcome these limitations, such as using multiple GPUs or more efficient memory management techniques, have shown promise in recent research.

Recent works have focused on optimizing the performance of ARM algorithms on GPUs by addressing the memory and communication bottlenecks inherent in large-scale data mining. For instance, Jiayi et al.[26] They introduced the Mempack strategy for compressing itemsets and reducing memory fetching operations during ARM on GPUs. While this approach showed speedups of up to 15x, it was limited by the complexity of the data structures involved. Other approaches have explored clustering techniques, such as the probabilistic frequent itemset mining approach.[27], which partitioned the dataset across GPU clusters. However, load balancing between nodes remained a challenge, impacting overall performance.

The introduction of swarm intelligence algorithms, such as Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO), has also been adapted for GPU architectures. These algorithms exploit the multi-threading capabilities of GPUs, allowing large numbers of swarm agents to explore the solution space concurrently. Weiss et al. implemented ANtMinerGPU.[28], an ACO algorithm where each ant is mapped to a GPU thread, demonstrating that larger swarms significantly improved solution quality. Similarly, Zhou et al.[29] We have developed a fully GPU-implemented PSO algorithm, where each dimension of the fitness function is evaluated in parallel, showing that the size of the swarm plays a crucial role in the parallel efficiency of the algorithmS. In the work by Cecilia et al.[30], they applied this hierarchical approach to solving the Traveling Salesman Problem (TSP) using a GPU-based ACO algorithm. In this setup, two categories of artificial ants were defined: queen ants and worker ants. The queen ants are associated with CUDA thread blocks, and the worker ants are individual CUDA threads that assist the queen ants in making decisions during the tour construction step of the algorithm. This collaboration between queen ants and worker ants helps accelerate the decision-making process, especially in selecting the next city to visit in the TSP. By using this strategy, the algorithm achieves better GPU resource utilization without requiring an excessively large swarm, which is beneficial for improving both performance and efficiency.

This hierarchical organization contrasts with the flat implementations, like those used in ANtMinerGPU , where each thread is assigned a single agent, potentially leading to under-utilization of the GPU's computational power if the swarm size is not large enough. By organizing threads hierarchically, the computational load is more evenly distributed, resulting in higher GPU occupancy and faster execution, particularly for complex tasks like ACO.

In [31], in parallel Association rule mining using GPUs and Bees behaviors, the authors propose a parallel GPU-based extension of the Bees Swarm Optimization algorithm for Association Rule Mining. While the CPU handles the solution generation and search processes, the evaluation of solutions is parallelized on the GPU to take advantage of its massive threading capabilities. The experimental results demonstrate that this parallel approach significantly outperforms the sequential version in terms of efficiency and execution time, showcasing the potential of GPU-based optimization in ARM.

In GPU-based Bees Swarm Optimization for Association Rules Mining[32], the paper introduces two GPU-based Bees Swarm Optimization algorithms for Association Rule Mining: SE-GPU and ME-GPU. SE-GPU evaluates one rule at a time, assigning each thread to a transaction, while ME-GPU evaluates multiple rules concurrently, with each thread handling several transactions. These approaches leverage

the parallel processing power of GPUs to address the computational challenges posed by large datasets, such as the WebDocs benchmark. Experiments conducted on an Intel Xeon processor coupled with an Nvidia Tesla GPU demonstrate up to 100× speed improvement compared to sequential versions. Additionally, the GPU-based algorithms outperformed multi-core CPU versions, proving more efficient across various core counts. This study highlights the significant advantages of GPU-based parallelism for large-scale ARM.

In 2020[33], researchers introduced two parallel Genetic Algorithms for ARM: ARM-GPU and ARM-CPU/GPU. ARM-GPU accelerates fitness evaluation using GPU parallelism, while ARM-CPU/GPU applies a two-level parallelism where CPU cores evolve sub-populations and offload fitness computation to the GPU. These methods significantly outperform traditional exact algorithms (Apriori, FP-Growth) and other parallel approaches (SE-GPU, ME-GPU) in terms of execution time, especially for large datasets. This demonstrates the effectiveness of combining CPU and GPU parallelism to handle the computational demands of large-scale ARM tasks.

In [34], the work introduced NRFP-Growth, a non-recursive version of FP-Growth, and GPFP-Growth, a parallel GPU-based variant. NRFP-Growth reduces time and space complexity by using FP-array and ItemPoss-map structures, while GPFP-Growth leverages GPU parallelism to accelerate frequent itemset mining. Tests on various datasets showed that NRFP-Growth improved performance over the classical FP-Growth, and GPFP-Growth provided significant speedups and scalability, making it suitable for large datasets.

Despite these advances, several challenges remain. Memory limitations on GPUs continue to constrain the size of datasets that can be processed efficiently. Additionally, the overhead of frequent data transfers between the CPU and GPU limits the potential speedups in some applications. Researchers are now investigating strategies to minimize these communication costs, such as by performing most computations on the GPU and reducing the need for CPU intervention.

3 Reptile Search Algorithm for ARM

The Reptile Search Algorithm (RSA) was introduced by Abualigah et al. in 2022 as a novel metaheuristic inspired by the hunting strategies of reptiles, particularly their encircling and cooperation behaviors. RSA has shown remarkable success in solving various NP-complete optimization problems, ranging from energy optimization to image processing.

Recently, RSA was applied to ARM by Boukhalat et al. in their work titled "Reptile Search Algorithm for Association Rule Mining." [7]. Their approach demonstrated that RSA could outperform traditional metaheuristics, such as BSO and genetic algorithms, in terms of rule quality, execution time, and memory efficiency. This application of RSA to ARM introduced a novel way to address the computational complexity of mining association rules from large datasets.

In the context of ARM, RSA is used to find the optimal association rules by evaluating candidate rules based on two key metrics:

- **Support:** The support of an association rule $X \rightarrow Y$ is determined by the frequency of occurrence of a rule's antecedent and consequent in the dataset, as shown in Eq (1).

$$\text{Support}(X \rightarrow Y) = \frac{P(X \cup Y)}{|D|}$$

- **Confidence:** The confidence (Conf) for the rule $X \rightarrow Y$, as outlined in Eq. (2), The likelihood that the consequent of the rule occurs given the antecedent.

$$\text{Confidence}(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X)}$$

The core components of RSA in ARM are:

- **Rule Representation:** Each rule is represented as a candidate solution in RSA. The rule's antecedent and consequent form the search space, which RSA explores.
- **Fitness Function:** The fitness of each rule is evaluated based on the support and confidence metrics as in Eq (3). The goal is to maximize both metrics to generate high-quality association rules.

$$F(r) = \begin{cases} \alpha * \text{Sup}(r) + \beta * \text{Conf}(r) & \text{if accepted} \\ -1 & \text{otherwise} \end{cases}$$

4 Proposed parallel approaches

4.1 Motivation for GPU-Accelerated RSA-ARM

Despite the success of metaheuristic algorithms like our previous RSA-ARM, as shown in Algorithm 1[7], the scalability of these algorithms remains a challenge, particularly when applied to large datasets. To address this, researchers have turned to parallel computing, utilizing multi-core CPUs and GPUs to reduce execution times. Graphics Processing Units (GPUs), with their high degree of parallelism, are particularly well-suited for tasks that involve repetitive computations, such as the evaluation of candidate rules in ARM.

Several studies have demonstrated the effectiveness of using GPUs to parallelize ARM algorithms. For instance, Djenouri et al. proposed a GPU-based implementation of BSO-ARM, which showed significant performance improvements over CPU-based implementations. Similarly, Syed and Qamar [35] demonstrated the effectiveness of parallelizing the Apriori algorithm using CUDA to achieve substantial speedups in support counting. However, no previous work has applied GPU parallelism to RSA for ARM, leaving an important gap in the literature for further performance optimization.

The fitness function in RSA-ARM, which evaluates the quality of association rules based on support and confidence, is computationally expensive. With larger datasets, the iterative nature of RSA further increases the complexity, making it difficult to achieve real-time performance using traditional CPU-based implementations. Therefore, by offloading this computationally intensive part of the algorithm to a GPU, the evaluation of candidate rules can be parallelized, resulting in a significant reduction in execution time.

Our work seeks to address this gap by implementing RSA-ARM on GPUs, parallelizing the evaluation of candidate rules and fitness functions to enhance scalability and performance on large datasets.

4.2 Parallelizing RSA-ARM Using the Master-Slave Paradigm

Parallelizing the RSA-ARM algorithm is based on the Master-Slave paradigm, also known as Master-Worker or Manager-Worker.[36]. This is a common parallel computing design pattern used in distributed systems to divide computational tasks across processors effectively. In this strategy, the CPU acts as the master, while the GPU functions as the slave.

The Master (CPU) coordinates the entire computational process, managing the generation of candidate solutions (rules represented by reptiles), distributing computationally intensive tasks (like fitness evaluations) to the GPU, and updating global information such as the best fitness or solution based on results received from the GPU. The Slaves (GPU) handle the execution of these computationally intensive tasks, distributing the workload across thousands of GPU cores. Each core performs computations in parallel, significantly speeding up the overall process.

4.3 Implementation in RSA-ARM

The CPU(Master) generates candidate solutions represented as reptiles in RSA-ARM. It manages the iterative flow of the algorithm, offloading computationally heavy parts (like fitness evaluation) to the GPU for parallel execution. Once results are received from the GPU, the CPU updates global information, such as the best fitness or solution found so far.

The GPU(Slave) performs the actual computations, including evaluating the support and confidence of candidate rules. The computational workload is distributed across the thousands of GPU cores, ensuring efficient parallel execution. In RSA-ARM, tasks such as fitness function evaluation and rule updates (high walking, belly walking, and hunting cooperation) are executed concurrently by the GPU cores.

As shown in Figure 1, the CPU (master) generates candidate solutions (rules represented by reptiles) and manages the iterative flow of the algorithm. It sends computationally expensive parts, such as fitness evaluations, to the GPU (slave) for parallel execution. After receiving results from the GPU, the CPU updates global information, such as the best fitness or solution. The GPU (slave) distributes the computational workload across its cores, performing calculations in parallel to optimize performance.

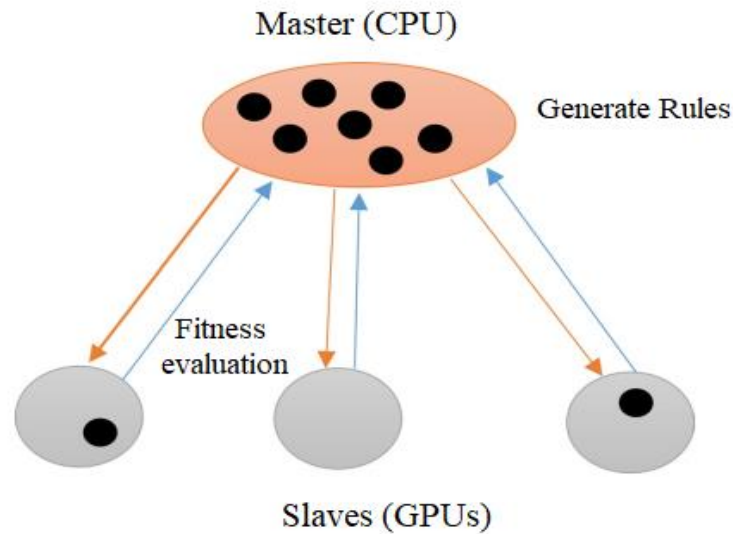


Figure 1. Master/Slave paradigm

4.4 Parallel Fitness Evaluation on GPU

Algorithm 2 shows that the fitness function evaluation is parallelized using the GPU to enhance the computational efficiency of RSA-ARM. At the start of the computation, the transaction data is transferred to the GPU memory to minimize overhead. This data remains in GPU memory throughout the process, ensuring minimal data transfer between the CPU and GPU. The dataset is split among multiple GPU threads, with each thread processing a portion of the dataset to calculate the support and confidence for different candidate rules. Each thread runs the fitnessKernel (Algorithm 3), which

computes support and confidence for a candidate rule, and the results are aggregated across threads, enabling faster fitness evaluations compared to a serial CPU implementation. The GPU's parallelization distributes fitness calculations across thousands of cores, significantly reducing execution time compared to CPU-only implementations.

4.5 Parallel Reptile Movements on GPU

In RSA-ARM, reptiles explore and exploit the search space using specific strategies like high walking, belly walking, and hunting cooperation. These strategies involve updating candidate solutions (rules) during each iteration. The movement strategies of RSA are parallelized by assigning each GPU thread to a reptile agent, ensuring that rule updates are performed concurrently. A separate CUDA kernel is designed to handle these reptile movements, with each thread applying the rule update logic in parallel, making the algorithm more scalable and efficient. In RSA-ARM, parallelization of rule updates ensures that computationally intensive parts of the algorithm are efficiently handled by the GPU, allowing concurrent updates of multiple candidate solutions.

Algorithm 1 Reptile Search Algorithm for ARM

Require: Max number of iterations T , Reptile population, Min support, and Min confidence

- 1: Initialize RSA criterion α , β , etc.
- 2: Initialize randomly the solutions X : $i = 1, \dots, N$.
- 3: **while** $v < T$ **do**
- 4: Compute the Fitness for the nominee solutions (X).
- 5: Determine the current best solution.
- 6: Modify the ES using Equations (8).
- 7: The start of the RSA.
- 8: **for** $i = 1$ to N **do**
- 9: Updating and adjusting the UB and LB.
- 10: Generate a new solution X_i using Algorithm 3.
- 11: **for** $j = 1$ to n **do**
- 12: Update the R, P , and eta using Equations (6), (7), and (9), respectively.
- 13: **if** $(v \leq \frac{T}{4})$ **then**
- 14: $x_{i,j}(1 + v) = [\beta \times -\eta_{i,j}(v) \times Best_j(v)] - \text{rand} \times R_{i,j}(v)$ // High walking
- 15: **else if** $(v \leq \frac{2T}{4} \text{ and } v > \frac{T}{4})$ **then**
- 16: $x_{i,j}(1 + v) = ES(v) \times Best_j(v) \times \text{rand} \times x_{r1,j}$ // Belly walking
- 17: **else if** $(v \leq \frac{3T}{4} \text{ and } v > \frac{2T}{4})$ **then**
- 18: $x_{i,j}(1 + v) = P_{i,j}(v) \times \text{rand} \times Best_j(v)$ // Hunting coordination
- 19: **else**
- 20: $x_{i,j}(1 + v) = Best_j(v) - \eta_{i,j}(v) \times \epsilon - \text{rand} \times R_{i,j}(v)$ // Hunting cooperation
- 21: **end if**
- 22: **end for**
- 23: **end for**
- 24: Calculate the new Maxfitness value for each Reptile agent.
- 25: Rank Reptile to the best solution.
- 26: $v = v + 1$
- 27: **end while**
- 28: **Output:** Best solution found, Fitness values for each reptile agent, and set of rules
- 29: Post-process results and visualization.

Figure 2. RSA-ARM[7]

Algorithm 2 Parallel RSA-ARM Using the Master-Slave Paradigm

Require: Dataset D , Number of reptiles R , Min support minsup, Min confidence minconf, Max number of generations G , GPU configuration with T threads

- 1: Initialize RSA parameters $\alpha, \beta, \alpha_{\text{Reptile}}, \beta_{\text{Reptile}}$
- 2: Generate initial set of candidate solutions (reptiles) X_i , for $i = 1, 2, \dots, R$
- 3: Transfer dataset D to GPU memory
- 4: Set generation $g = 0$
- 5: **while** $g < G$ **do**
- 6: **Master (CPU) Coordination**
- 7: **for** each reptile $i = 1, 2, \dots, R$ **do**
- 8: Generate candidate solutions and manage updates using movement strategies
- 9: Offload computationally heavy tasks (fitness evaluation) to GPU
- 10: **end for**
- 11: **Parallel Fitness Evaluation on GPU**
- 12: **for** each GPU thread $t = 1, 2, \dots, T$ **do**
- 13: Split dataset D among T GPU threads
- 14: Calculate support and confidence for each rule
- 15: **if** support $\geq$ minsup **and** confidence $\geq$ minconf **then**
- 16: Compute fitness: $\text{Fitness}(r) = \beta \times \text{support}(r) + \alpha \times \text{confidence}(r)$
- 17: **end if**
- 18: **end for**
- 19: Aggregate fitness results across GPU threads and return to CPU
- 20: **Master (CPU) Updates**
- 21: **for** each reptile $i = 1, 2, \dots, R$ **do**
- 22: Update global best fitness value and solution
- 23: **for** each attribute $j = 1, 2, \dots, n$ **do**
- 24: Update reptile positions based on movement strategies:
- 25: **if** $g \leq \frac{G}{4}$ **then**
- 26: High Walking: $X_i[j] = X_i[j] - \eta_i \times \beta_{\text{Reptile}} - R_i[j] \times \text{rand}()$
- 27: **else if** $\frac{G}{4} < g \leq \frac{2G}{4}$ **then**
- 28: Belly Walking: $X_i[j] = X_i[j] \times X_{\text{random}}[j] \times \text{ES} \times \text{rand}()$
- 29: **else if** $\frac{2G}{4} < g \leq \frac{3G}{4}$ **then**
- 30: Hunting Coordination: $X_i[j] = X_i[j] \times P_i[j] \times \text{rand}()$
- 31: **else**
- 32: Hunting Cooperation: $X_i[j] = X_i[j] - \eta_i \times \text{Double.MIN} - R_i[j] \times \text{rand}()$
- 33: **end if**
- 34: **end for**
- 35: **end for**
- 36: **Reptile Movement Parallelization on GPU**
- 37: **for** each GPU thread $t = 1, 2, \dots, T$ **do**
- 38: Assign reptile agent to thread and apply movement strategies concurrently
- 39: **end for**
- 40: Update reptile solutions and increment generation: $g = g + 1$
- 41: **end while**
- 42: **Output:** Best solution found, Fitness values for each reptile agent, and set of association rules
- 43: Post-process results and visualization

Figure 3. Parallel RSA-ARM Algorithm

Algorithm 3 Fitness Calculation for Parallel RSA-ARM

Require:

- lhs*: Array representing the left-hand side items of the rule
- lhsSize*: Number of items in *lhs*
- 3: *rhs*: Array representing the right-hand side items of the rule
- rhsSize*: Number of items in *rhs*
- α, β : Parameters for the fitness calculation
- 6: minsup, minconf: Minimum support and confidence thresholds

Ensure:

result: The fitness value for the candidate rule

Steps:**Thread ID Calculation:**

- 9: Calculate the thread ID as:

$$\text{tid} = \text{blockIdx.x} \times \text{blockDim.x} + \text{threadIdx.x}$$

This calculation assigns each GPU thread to process one item in the *lhs* array.

Check Bounds:

- 12: **if** tid $\geq$ *lhsSize* **then**
 Terminate the current thread
else
15: Proceed to the next step
end if

Initialize Support and Confidence:

- 18: Set *support* $\leftarrow$ 0.5 $\triangleright$ Example value for support calculation
 Set *confidence* $\leftarrow$ 0.6 $\triangleright$ Example value for confidence calculation

Check Minimum Thresholds:

- 21: **if** *support* $\geq$ minsup **and** *confidence* $\geq$ minconf **then**
 Proceed to the next step
else
24: Set *fitness* $\leftarrow$ -1 and terminate the current thread
end if

Calculate Fitness:

- 27: Compute the fitness of the candidate rule as:

$$\text{Fitness} = \beta \times \text{support} + \alpha \times \text{confidence}$$

Atomic Update of Result:

Use an atomic addition operation to safely update the global *result*:

$$\text{result} \leftarrow \text{result} + \text{fitness}$$

- 30: This ensures that concurrent threads do not cause race conditions when updating the shared variable.

Figure 4. Fitness calculation for parallel RSA-ARM

5 Performance Evaluation

In this section, we first evaluate the performance of the proposed Parallel RSA-ARM using GPUs against the original CPU-based RSA-ARM. In the second part, we will make comparisons with state-of-the-art GPU-accelerated ARM algorithms, such as Multiple Evaluation GPU-based BSO. The comparison was conducted using the same datasets to ensure a fair and accurate evaluation.

5.1 Experimental Setup

Table 1 shows hardware specifications and software framework environment. The experiments were conducted on a system configured as follows:

Table 1: Hardware specifications and software framework environment

Category	Specification
Hardware Specifications	CPU: Intel Core i7 Cores/Threads: 8 cores, 8 threads 1. RAM: 16 GB DDR4 GPU: NVIDIA GeForce RTX 2070, 8 GB
Software Frameworks	CUDA Toolkit 11.4: Libraries and tools for parallel execution on NVIDIA GPUs. JCuda Library: Java binding for CUDA enabling Java applications to interact with CUDA. Java Development Kit (JDK) 11: For implementing RSA-ARM algorithm (both CPU and GPU versions).

5.2 Datasets and Performance Metrics

The evaluation uses the same benchmark datasets from the original CPU-based RSA-ARM experiments. The datasets represent varying sizes and complexities, allowing us to measure performance across small, medium, and large datasets. To compare the CPU-based and GPU-based RSA-ARM implementations, metrics such as Execution Time, Speedup Ratio, which is the ratio of execution times between the CPU and GPU implementations, and scalability are considered.

5.3 Results

The execution times and Speedup Ratio of the CPU-based RSA-ARM, Parallel RSA-ARM ME-GPU-BSO for each dataset are presented in Table 2

Table 2: CPU-based RSA-ARM vs Parallel RSA-ARM vs ME-GPU-BSO

Dataset	Transactions	Item Size	CPU-based RSA-ARM Time (s)	GPU-based RSA-ARM Time (s)	ME-GPU-BSO Time (s)	Speed up Ratio
Basketball	96	5	0.17	0.04	0.69	4.25
IBM Quest	1,000	20	0.13	0.03	1.73	4.33
Quake	2,178	4	0.10	0.02	1.38	5.00
Chess	3,196	75	0.30	0.08	16.38	3.75
Mushroom	8,124	119	1.42	0.35	25.86	4.06
BMS-Web-1	59,602	497	40.78	8.5	106	3.88
BMS-POS	515,597	1,657	1,185	125	2166	9.48
WebDocs	1,692,082	526,765	Blocked	1,500	6202	/

5.4 Discussion

5.4.1 Comparing with CPU-based RSA-ARM

The Parallel GPU-based RSA-ARM consistently outperforms the original CPU-based RSA-ARM across all datasets, demonstrating clear performance improvements due to parallelization. The speedup ratios range from 3.75x to 5.00x, which aligns with common speedup trends in GPU-accelerated data mining tasks. This indicates that the algorithm is effectively leveraging the parallel architecture of the GPU. The results validate the potential of parallel computing to handle the computational bottlenecks in RSA-ARM, especially in evaluating the fitness function across large datasets. For smaller datasets like Basketball, IBM Quest, and Quake, the GPU implementation achieves speedup ratios of around 4x to 5x. For medium-sized datasets like Chess and Mushroom, the speedup ratios are slightly lower, around 3.75x to 4.06x.

For medium-sized datasets like Chess and Mushroom, the speedup ratios are slightly lower, around 3.75x to 4.06x. This is due to increased memory overhead and more complex rule mining, which requires more GPU resources but still maintains significant speedup compared to the CPU. In large datasets like BMS-Web-1 and BMS-POS, the GPU-based RSA-ARM achieves speedup ratios of approximately 9.48x, which reflects the ability of the GPU to manage larger transaction volumes and item sizes through efficient parallel computation.

The execution time for BMS-POS is reduced from 1185 seconds on the CPU to 125 seconds on the GPU, demonstrating a substantial reduction in mining time. The CPU-based implementation could not process the WebDocs dataset due to memory limitations, but it was feasible on the GPU within 1,500 seconds. This shows that parallelization not only accelerates computation but also enables the processing of previously intractable datasets.

The results demonstrate the scalability of the GPU-based RSA-ARM, as it maintains consistent speedup ratios across datasets of varying sizes. However, there is a slight reduction in speedup for larger datasets due to Handling larger transaction data and frequent itemsets, which can lead to more memory access delays, impacting speedup, and transferring data between CPU and GPU can add overhead, especially for extremely large datasets. Despite these limitations, the Parallel RSA-ARM shows strong potential to efficiently mine association rules even in complex, large-scale datasets.

5.4.2 Comparing with other GPU-accelerated approaches

GPU-RSA-ARM is extremely fast for small datasets; ME-GPU-BSO lags significantly, with runtimes 17–69 times slower than GPU-RSA-ARM. As the dataset size increases, GPU-RSA-ARM maintains its performance with minimal increases in runtime. ME-GPU-BSO slows down significantly, with runtimes 74–204 times slower, demonstrating poor scalability compared to GPU-RSA-ARM.

GPU-RSA-ARM scales efficiently to large datasets, maintaining reasonable runtimes even for extremely complex datasets like WebDocs. ME-GPU-BSO, while feasible, becomes 4–17 times slower, with runtimes that are impractical for real-time or large-scale applications.

6 Conclusion

The results strongly support the use of GPU-based parallelization for RSA-ARM, showing significant performance gains across all dataset sizes. The approach not only accelerates mining but also extends the feasibility of processing larger datasets that are beyond the capacity of traditional CPU-based implementations. This work demonstrates that parallel computing is a promising strategy for enhancing the speed and scalability of association rule mining, making it more applicable to big data scenarios. These findings illustrate the effectiveness of the Parallel RSA-ARM and highlight its advantages in terms of speed, scalability, and handling of complex datasets.

References

- [1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. Very large databases, VLDB*, Citeseer, 1994, pp. 487–499.
- [2] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1–12, Jun. 2000, doi: 10.1145/335191.335372.
- [3] Y. Djenouri, H. Drias, and Z. Habbas, "Bees swarm optimisation using multiple strategies for association rule mining," *International Journal of Bio-Inspired Computation*, vol. 6, no. 4, p. 239, 2014, doi: 10.1504/IJBIC.2014.064990.
- [4] Z. Kou and L. Xi, "Binary Particle Swarm Optimization-Based Association Rule Mining for Discovering Relationships between Machine Capabilities and Product Features," *Math Probl Eng*, vol. 2018, pp. 1–16, Oct. 2018, doi: 10.1155/2018/2456010.
- [5] M. M. Dorigo Vittorio; Colorni Alberto, "Ant system: optimization by a colony of cooperating agents," *IEEE Trans Syst Man Cybern B Cybern*, vol. 26, no. 1, pp. 29–41, 1996, doi: 10.1109/3477.484436.
- [6] L. Abualigah, M. A. Elaziz, P. Sumari, Z. W. Geem, and A. H. Gandomi, "Reptile Search Algorithm (RSA): A nature-inspired meta-heuristic optimizer," *Expert Syst Appl*, vol. 191, p. 116158, Apr. 2022, doi: 10.1016/j.eswa.2021.116158.
- [7] A. Boukhalat, K. E. Heraguemi, M. Benouis, B. Bouderah, and S. Akhrouf, "Reptile Search Algorithm for Association Rule Mining," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1729–1744, 2024, doi: 10.12785/ijcds/1501122.
- [8] S. S. Aljehani and Y. A. Alotaibi, "Preserving Privacy in Association Rule Mining Using Metaheuristic-Based Algorithms: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 21217–21236, 2024, doi: 10.1109/ACCESS.2024.3362907.
- [9] Y. Zhang *et al.*, "Parallel Processing Systems for Big Data: A Survey," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2114–2136, Nov. 2016, doi: 10.1109/JPROC.2016.2591592.
- [10] M. Madijagan and S. S. Raj, "Chapter 1 - Parallel Computing, Graphics Processing Unit (GPU) and New Hardware for Deep Learning in Computational Intelligence Research," in *Deep Learning and Parallel Computing Environment for Bioengineering Systems*, A. K. Sangaiah, Ed., Academic Press, 2019, pp. 1–15. doi <https://doi.org/10.1016/B978-0-12-816718-2.00008-7>.

-
- [11] H. Bavarsad Salehpour, H. Haj Seyyed Javadi, P. Asghari, and M. E. Shiri Ahmad Abadi, "Improvement of Apriori Algorithm Using Parallelization Technique on Multi-CPU and GPU Topology," *Wirel Commun Mob Comput*, vol. 2024, pp. 1–14, May 2024, doi: 10.1155/2024/7716976.
 - [12] S. H. Adil and S. Qamar, "Implementation of association rule mining using CUDA," in *2009 International Conference on Emerging Technologies*, IEEE, Oct. 2009, pp. 332–336. doi: 10.1109/ICET.2009.5353149.
 - [13] M. Kaya and R. Alhajj, "Utilizing Genetic Algorithms to Optimize Membership Functions for Fuzzy Weighted Association Rules Mining," *Applied Intelligence*, vol. 24, no. 1, pp. 7–15, Feb. 2006, doi: 10.1007/s10489-006-6925-0.
 - [14] C. A. C. Coello and M. S. Lechuga, "MOPSO: A proposal for multiple objective particle swarm optimization," in *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, IEEE, 2002, pp. 1051–1056.
 - [15] Y. Djenouri, H. Drias, and Z. Habbas, "Bees swarm optimisation using multiple strategies for association rule mining," *International Journal of Bio-Inspired Computation*, vol. 6, no. 4, p. 239, 2014, doi: 10.1504/IJBIC.2014.064990.
 - [16] R. Agrawal and J. C. Shafer, "Parallel mining of association rules," *IEEE Trans Knowl Data Eng*, vol. 8, no. 6, pp. 962–969, 1996, doi: 10.1109/69.553164.
 - [17] D. Boley *et al.*, "Partitioning-based clustering for Web document categorization," *Decis Support Syst*, vol. 27, no. 3, pp. 329–341, Dec. 1999, doi: 10.1016/S0167-9236(99)00055-X.
 - [18] B. K. Sarker, T. Mori, T. Hirata, and K. Uehara, "Parallel Algorithms for Mining Association Rules in Time Series Data," 2003, pp. 273–284. doi: 10.1007/3-540-37619-4_28.
 - [19] O. R. Zaiane, M. El-Hajj, and P. Lu, "Fast parallel association rule mining without candidacy generation," in *Proceedings 2001 IEEE International Conference on Data Mining*, IEEE Comput. Soc, pp. 665–668. doi: 10.1109/ICDM.2001.989600.
 - [20] N. Melab and E. Talbi, "A parallel genetic algorithm for rule mining," in *Proceedings 15th International Parallel and Distributed Processing Symposium. IPDPS 2001*, IEEE Comput. Soc, pp. 1347–1352. doi: 10.1109/IPDPS.2001.925112.
 - [21] J. Eggermont, J. N. Kok, and W. A. Kusters, "Genetic Programming for data classification," in *Proceedings of the 2004 ACM symposium on Applied computing*, New York, NY, USA: ACM, Mar. 2004, pp. 1001–1005. doi: 10.1145/967900.968104.
 - [22] G. Agapito, P. H. Guzzi, and M. Cannataro, "Parallel and distributed association rule mining in life science: A novel parallel algorithm to mine genomics data," *Inf Sci (N Y)*, vol. 575, pp. 747–761, Oct. 2021, doi: 10.1016/j.ins.2018.07.055.
 - [23] W. Fang, M. Lu, X. Xiao, B. He, and Q. Luo, "Frequent itemset mining on graphics processors," in *Proceedings of the Fifth International Workshop on Data Management on New Hardware*, New York, NY, USA: ACM, Jun. 2009, pp. 34–42. doi: 10.1145/1565694.1565702.
-

- [24] K.-W. Chon, S.-H. Hwang, and M.-S. Kim, "GMiner: A fast GPU-based frequent itemset mining method for large-scale data," *Inf Sci (N Y)*, vol. 439–440, pp. 19–38, May 2018, doi: 10.1016/j.ins.2018.01.046.
- [25] A. Cano, J. M. Luna, and S. Ventura, "High-performance evaluation of evolutionary-mined association rules on GPUs," *J Supercomput*, vol. 66, no. 3, pp. 1438–1461, Dec. 2013, doi: 10.1007/s11227-013-0937-4.
- [26] X. J. A. Bellekens, C. Tachtatzis, R. C. Atkinson, C. Renfrew, and T. Kirkham, "A Highly-Efficient Memory-Compression Scheme for GPU-Accelerated Intrusion Detection Systems," in *Proceedings of the 7th International Conference on Security of Information and Networks*, New York, NY, USA: ACM, Sep. 2014, pp. 302–309. doi: 10.1145/2659651.2659723.
- [27] Y. Djenouri, A. Belhadi, P. Fournier-Viger, and H. Fujita, "Mining diversified association rules in big datasets: A cluster/GPU/genetic approach," *Inf Sci (N Y)*, vol. 459, pp. 117–134, Aug. 2018, doi: 10.1016/j.ins.2018.05.031.
- [28] R. Skinderowicz, "The GPU-based parallel Ant Colony System," *J Parallel Distrib Comput*, vol. 98, pp. 48–60, Dec. 2016, doi: 10.1016/j.jpdc.2016.04.014.
- [29] Y. Zhuo, T. Zhang, F. Du, and R. Liu, "A parallel particle swarm optimization algorithm based on GPU/CUDA," *Appl Soft Comput*, vol. 144, p. 110499, Sep. 2023, doi: 10.1016/j.asoc.2023.110499.
- [30] X. Geng, Z. Chen, W. Yang, D. Shi, and K. Zhao, "Solving the traveling salesman problem based on an adaptive simulated annealing algorithm with greedy search," *Appl Soft Comput*, vol. 11, no. 4, pp. 3680–3689, Jun. 2011, doi: 10.1016/j.asoc.2011.01.039.
- [31] Y. Djenouri, A. Bendjoudi, M. Mehdi, N. Nouali-Taboudjemat, and Z. Habbas, "Parallel association rules mining using GPUS and bees behaviors," in *6th International Conference on Soft Computing and Pattern Recognition, SoCPaR 2014*, Institute of Electrical and Electronics Engineers Inc., Jan. 2014, pp. 401–405. doi: 10.1109/SOCPAR.2014.7008040.
- [32] Y. Djenouri, A. Bendjoudi, M. Mehdi, N. Nouali-Taboudjemat, and Z. Habbas, "GPU-based bees swarm optimization for association rules mining," *Journal of Supercomputing*, vol. 71, no. 4, pp. 1318–1344, Apr. 2015, doi: 10.1007/s11227-014-1366-8.
- [33] L. Hamdad, Z. Ournani, K. Benatchba, and A. Bendjoudi, "Two-level parallel CPU/GPU-based genetic algorithm for association rule mining," *International Journal of Computational Science and Engineering*, vol. 22, no. 2/3, p. 335, 2020, doi: 10.1504/IJCSE.2020.107366.
- [34] B. Silva, L. G. Lopes, and F. Mendonça, "Parallel GPU-Acceleration of Metaphorless Optimization Algorithms: Application for Solving Large-Scale Nonlinear Equation Systems," *Applied Sciences*, vol. 14, no. 12, p. 5349, Jun. 2024, doi: 10.3390/app14125349.
- [35] S. H. Adil and S. Qamar, "Implementation of association rule mining using CUDA," in *2009 International Conference on Emerging Technologies*, IEEE, Oct. 2009, pp. 332–336. doi: 10.1109/ICET.2009.5353149.

- [36] A. Shenneld, P. Fleming, J. Allan, and V. Kadiramanathan, “Optimisation of Maintenance Scheduling Strategies on the Grid,” in *2007 IEEE Symposium on Computational Intelligence in Scheduling*, IEEE, Apr. 2007, pp. 231–237. doi: 10.1109/SCIS.2007.367695.

A Comprehensive Evaluation of YOLOv9, YOLOv10, and YOLOv11 in Crowd Counting Applications

Merwan Grabsi^{1,2✉}, Kameleddine Heraguemi³

¹ Computer Science, University Mohamed Boudiaf, M'Sila, Algeria

² Laboratory of Informatics and applications of M'sila LIAM

Merwan.grabsi@univ-msila.dz

³ National School of Artificial Intelligence, Algiers, Algeria

Kameleddine.heraguemi@ensia.edu.dz

Abstract

Recent advances in deep learning have significantly advanced the development of real-time object detection systems, with the YOLO (You Only Look Once) series being notable for its efficiency and accuracy. As public safety in crowded environments increasingly relies on smart surveillance technologies, optimizing these algorithms for crowd counting has become a priority. This study offers a comprehensive comparison of the latest YOLO versions—YOLOv9, YOLOv10, and YOLOv11—specifically in the context of crowd counting. Using recognized datasets such as ShanghaiTech and SCUTHead, we evaluate these models based on accuracy, detection speed, and computational efficiency. The experiments are designed to reflect real-world scenarios, making the findings applicable to practical surveillance challenges. Preliminary results indicate notable differences among the versions, each exhibiting unique strengths suitable for specific applications. This paper outlines the evolutionary improvements in the YOLO series and aims to guide developers in selecting the most appropriate version for public safety applications. Our findings have the potential to influence future designs and optimizations of real-time detection systems, enhancing safety in high-density areas.

Keywords: YOLO, Object detection, Crowd counting, Deep learning, Computer vision, CNN

1. Introduction

The increasing demand for real-time monitoring in critical applications, such as public safety, smart cities, and event management, has made accurate and efficient crowd counting a vital task in computer vision. Robust models are required to process large-scale crowd data in real time to detect and count individuals effectively, even in complex environments. Traditionally, crowd counting has faced significant challenges, including issues like severe occlusion, scale variation, and complex motion patterns[1][2]. Occlusions and clutter often hinder accurate detection, while variations in scale, caused by perspective effects, add another layer of complexity to the analysis[3]. Moreover, sophisticated models are needed to capture both macro-level and micro-level dynamics in real-time scenarios, making this a computationally intensive task.

The YOLO (You Only Look Once) family of models has emerged as a powerful solution in the field of object detection and crowd counting, offering an exceptional balance between speed and accuracy. YOLO's single-stage detection framework allows it to detect objects in real-time by processing the entire image at once, compared to traditional methods like R-CNN[3], Fast R-CNN[4] and Faster R-CNN[5] which rely on region proposals followed by classification. Since its introduction, YOLO has undergone significant improvements, especially in terms of architectural designs, optimization objectives, data

augmentation strategies, and more, making it one of the most widely adopted models in various applications, including pedestrian detection, crowd monitoring and crowd counting [6][7].

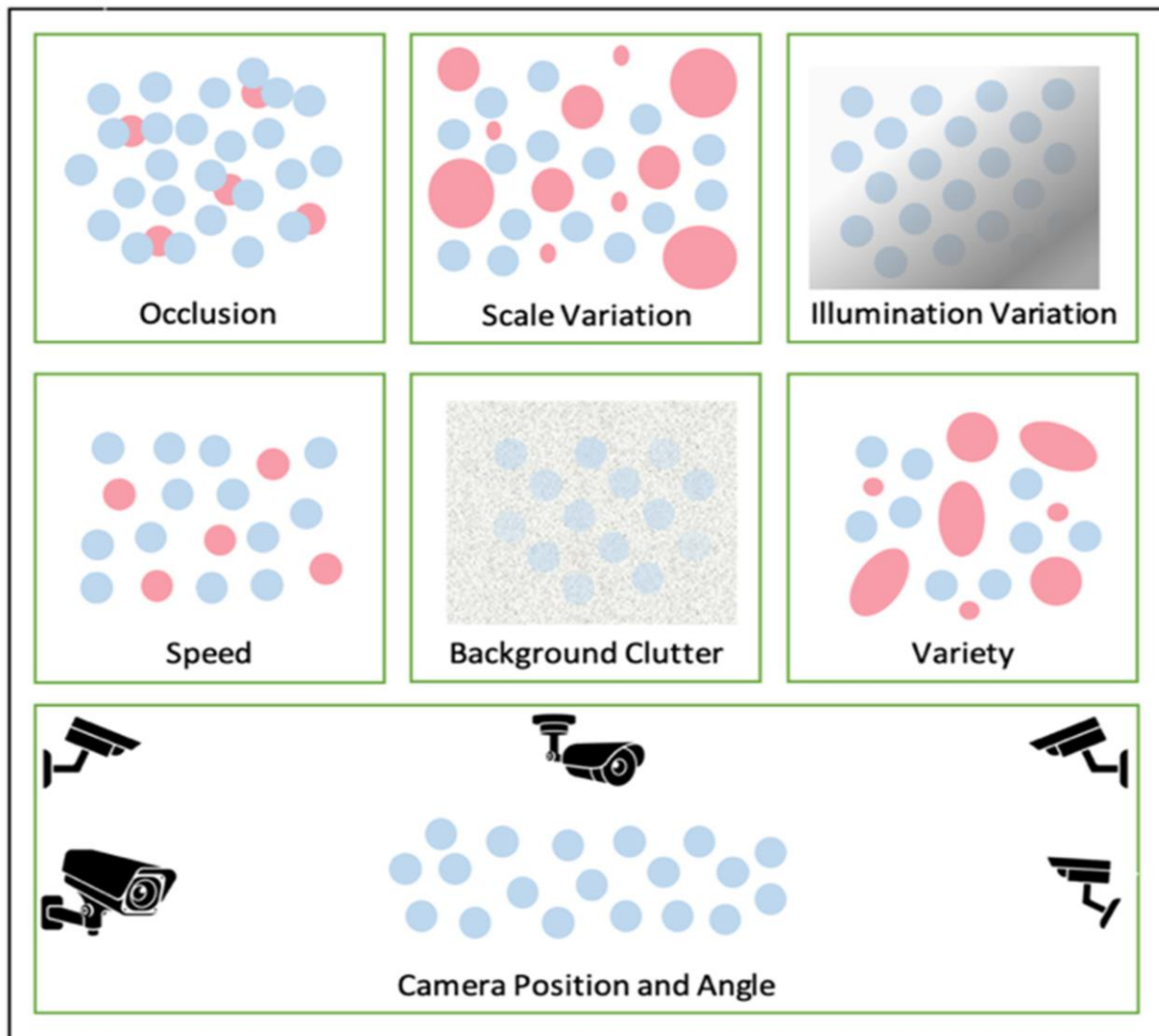


Figure 1. Crowd scene analysis challenges.[1]

Recent iterations of YOLO, particularly YOLOv9 [8], YOLOv10 [6], and the latest YOLOv11 [9], [10], have continued to push the boundaries of real-time object detection. YOLOv9 introduced significant innovations, such as Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN), which optimize the flow of information through the network, leading to improved accuracy in complex scenes [8]. YOLOv10 further enhanced the efficiency of real-time detection by eliminating the need for Non-Maximum Suppression (NMS), a technique commonly used to filter overlapping bounding boxes, thereby reducing computational redundancy and improving inference speed[6]. The latest version, YOLOv11, developed by Ultralytics [11] (Ultralytics is a tech company specializing in AI and machine learning, known for creating YOLOv5, YOLOv8, and the latest version, YOLOv11, which are widely used for real-time object detection. They provide open-source tools and resources for developers and researchers to implement cutting-edge computer vision solutions.), is designed to enhance further the model's ability to handle dense crowds and occlusions, making it a suitable candidate for challenging crowd counting scenarios[9].

While several studies have evaluated the performance of earlier YOLO versions like YOLOv5 and YOLOv8 [11], against other models like Detectron2 [12], the focus has predominantly been on general object detection rather than crowd counting [13]. Moreover, with the growing complexity of modern urban environments, there is a pressing need for models that can maintain high accuracy while efficiently processing large volumes of data in real-time. Thus, a detailed evaluation of YOLOv9, YOLOv10, and YOLOv11 in the context of crowd counting is both timely and necessary to determine their suitability for real-world applications like urban surveillance and public safety systems.

This study aims to fill this research gap by providing a comprehensive comparison of YOLOv9 (see Figure 2), YOLOv10 (see Figure 3), and YOLOv11 (see Figure 4), specifically in their ability to detect and count individuals in crowd environments. The following sections outline the specific objectives that will guide this comparative analysis.

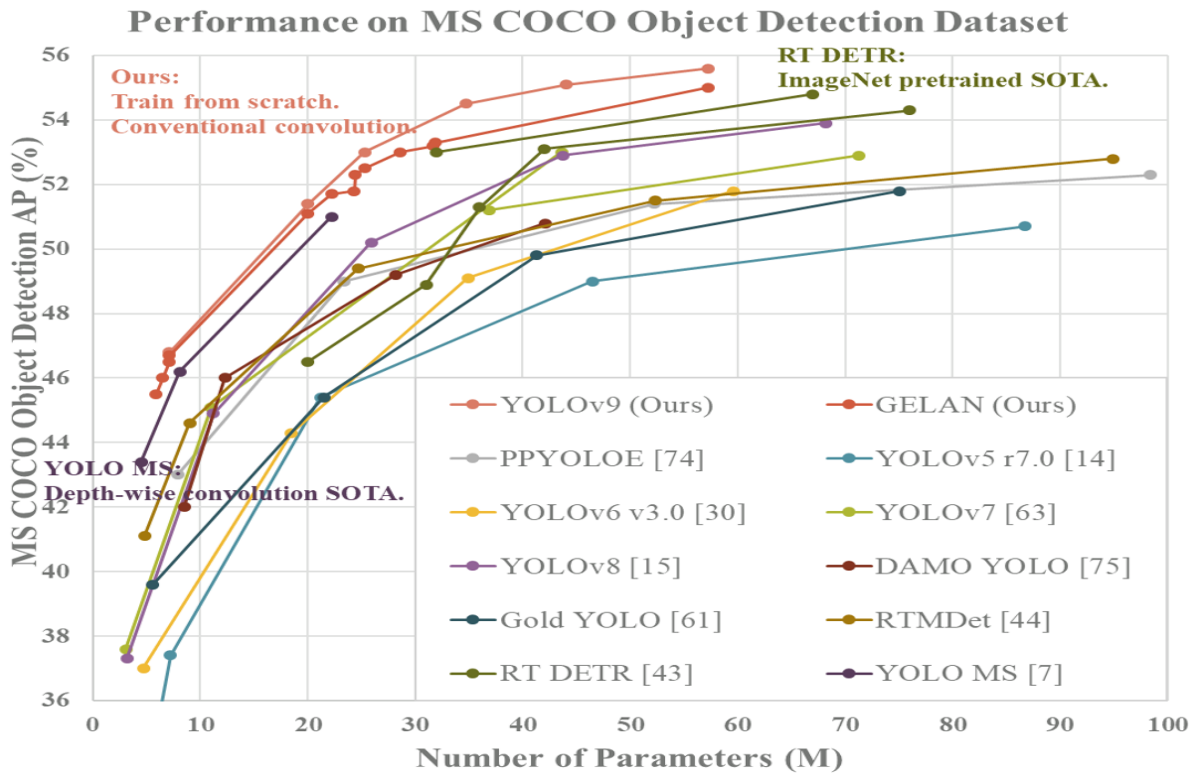


Figure 2. Performance Comparison of YOLOv9, and Other Object Detection Models on COCO[14] Dataset.

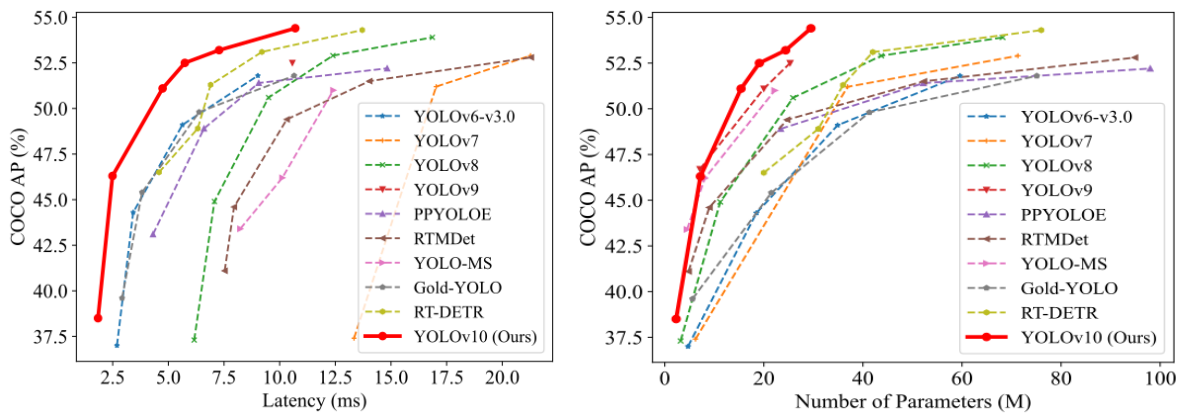


Figure 3. Performance and Efficiency Comparison of YOLOv10, and Other Object Detection Models on COCO [14] Dataset

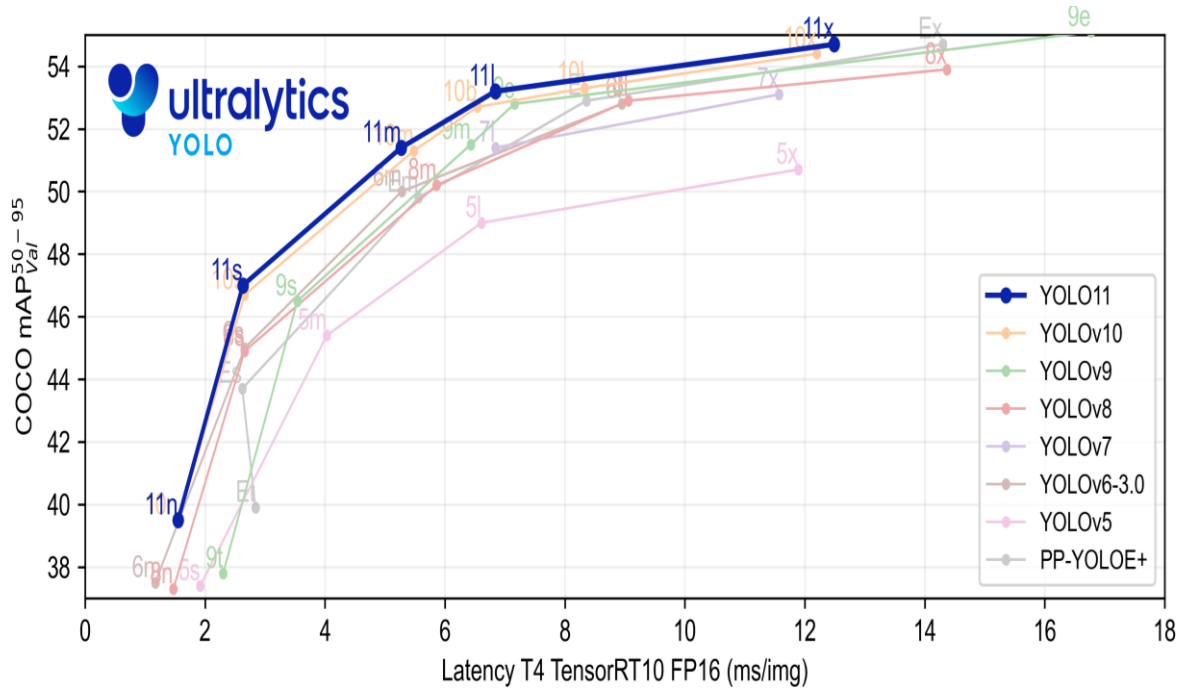


Figure 4. Performance Comparison of YOLOv11, and Other Object Detection Models on COCO[14] Dataset

1.2 Objectives

The primary objective of this study is to conduct a comprehensive performance analysis of the latest iterations of the YOLO (You Only Look Once) model—YOLOv9, YOLOv10, and YOLOv11—in the context of crowd counting. Specifically, the study aims to compare these models based on their ability to accurately detect and count individuals in crowded environments, with a focus on key metrics such as detection accuracy, speed, and computational efficiency.

1. **Evaluate Detection Accuracy and Counting Precision:** This objective focuses on assessing the detection accuracy of YOLOv9, YOLOv10, and YOLOv11 by measuring mean Average Precision (mAP) across high-density crowd environments. Additionally, the study will calculate crowd counting precision by evaluating Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) across different crowd densities.
2. **Measure Inference Speed and Latency:** It is crucial to measure the frames per second (FPS) and overall latency of each model in a crowd counting context. YOLOv9 and YOLOv10 introduced changes designed to improve inference speed, such as the removal of Non-Maximum Suppression (NMS)[15] in YOLOv10. YOLOv11 is expected to offer further speed optimizations for real-time deployment.
3. **Analyze Computational Resource Usage:** The study will examine the computational efficiency of the models, particularly their GPU and memory usage during the crowd counting task. This will help determine which model provides the best balance between accuracy and resource consumption.
4. **Assess Robustness in Congested and Complex Environments:** This objective will evaluate how well each model handles highly congested environments, where occlusions and overlapping objects pose significant challenges. The performance of YOLOv9, YOLOv10, and YOLOv11 will be compared to identify which model performs best in dense pedestrian traffic.
5. **Compare Performance Across Diverse Datasets:** Finally, the study will benchmark YOLOv9, YOLOv10, and YOLOv11 across multiple datasets, including Mall and ShanghaiTech, which are widely used for pedestrian detection and crowd counting.

2. Literature Review

The field of crowd counting has seen significant advancements with the emergence of deep learning models [16], particularly in object detection. Among the most prominent techniques are the YOLO (You Only Look Once) models, renowned for their capability to achieve real-time object detection while balancing speed and accuracy. This literature review presents an overview of key developments in YOLO-based models, explicitly focusing on YOLOv9, YOLOv10, and YOLOv11 and their applications in crowd counting.

2.1 YOLO Models for Object Detection and Crowd Counting

YOLO was initially proposed by Joseph Redmon in 2016 [2] and quickly gained popularity due to its single-stage object detection framework, which bypasses the need for region proposals, unlike the more traditional R-CNN family [3], [4], [5]. By framing object detection as a regression problem, YOLO can directly predict bounding boxes and class probabilities from the input image, resulting in much faster processing times compared to region-based models [7]. This speed advantage has made YOLO an attractive solution for real-time crowd counting in surveillance systems, event monitoring, and public safety.

With the release of YOLOv3 [17], significant architectural changes were made to improve accuracy, including the introduction of Darknet-53, a more powerful backbone network [13]. This version was applied successfully to various pedestrian detection and crowd counting tasks, with studies showing high accuracy rates on datasets such as INRIA[18] and ShanghaiTech[19] [20]. The use of YOLOv3 for pedestrian detection in both static images and video sequences demonstrated its effectiveness in real-world scenarios. For example, a study on pedestrian detection using YOLOv3 reported an accuracy of 96.1% on the INRIA dataset and 87.3% on the ShanghaiTech [19] dataset [20].

As the demand for more robust crowd counting models increased, YOLOv4[21] and YOLOv5[11] were introduced, focusing on improving model efficiency and performance in dense environments. YOLOv4 enhanced the feature extraction capabilities with Cross Stage Partial (CSP)[22] networks, while YOLOv5 brought significant improvements in both speed and accuracy, making it suitable for deployment in resource-constrained environments like edge computing.

2.2 Recent Advances: YOLOv9, YOLOv10, and YOLOv11

The introduction of YOLOv9 marked a shift towards more specialized architectures that optimized both the learning process and the model's ability to handle complex detection tasks. YOLOv9 introduced the concept of Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN), which allowed for more efficient parameter utilization during training [8], [23] . These innovations were particularly useful in handling dense and occluded crowd environments, where accurate detection and counting are challenging.

YOLOv10, building on the foundations laid by YOLOv9, introduced significant optimizations for real-time detection. One of the key innovations in YOLOv10 was the elimination of Non-Maximum Suppression (NMS), which is typically used to filter out overlapping bounding boxes in object detection tasks. By replacing NMS with a more efficient bounding box selection algorithm, YOLOv10 significantly reduced computational redundancy, making it faster without sacrificing accuracy. This made YOLOv10 particularly suited for high-density crowd counting scenarios, where the overlap between detected individuals can be substantial.

The most recent addition to the YOLO family, YOLOv11, was developed by Ultralytics and is designed specifically to enhance performance in challenging environments like congested crowds. YOLOv11 builds on the architectural improvements introduced in its predecessors, with a focus on optimizing real-

time performance while maintaining high detection accuracy. The integration of new techniques to handle occlusions and small object detection makes YOLOv11 a promising tool for crowd counting in environments like public events and transport hubs.

2.3 Comparative Studies and Performance Analysis

Several comparative studies have evaluated the performance of earlier YOLO versions (up to YOLOv8) in crowd counting tasks. For instance, a study comparing YOLOv8 and Detectron2 found that while YOLOv8 offered faster inference times, Detectron2 performed better in terms of detection accuracy[13]. This finding highlights the trade-off between speed and accuracy that is often encountered in real-time object detection models. However, YOLO models tend to be preferred for crowd counting due to their superior real-time capabilities.

The crowd counting performance of YOLO models has also been evaluated in various studies focusing on dense environments. For example, a modified version of YOLOv4 was developed to handle congested crowds by improving the detection of overlapping individuals, which is a common issue in such environments[24]. Similarly, studies using YOLOv3 for multi-object detection and counting demonstrated the model's efficiency in real-time scenarios, further cementing YOLO's role as a leading model for real-time applications[25].

A more recent analysis of one-stage object detection models[26], which includes YOLO, revealed that YOLOv7 [27] and YOLOv8 outperformed other single-stage models like (SSD: Single Shot MultiBox Detector[28]) when tested in dense environments. These findings further emphasize the strengths of YOLO models in environments where speed and efficiency are crucial.

2.4 Future Directions

As crowd counting becomes increasingly critical in urban environments, future research on YOLO models is expected to focus on enhancing their robustness in dynamic crowd scenarios, particularly for real-time video streams with fluctuating crowd densities [29]. Integrating YOLO models with behavioral analysis could provide valuable insights into crowd dynamics and safety. Additionally, optimizing YOLO models for edge AI will be a pivotal area of research, especially for applications in smart cities and public event monitoring.

3. Methodology

This section outlines the methodology adopted to rigorously evaluate the performance of YOLOv9, YOLOv10, and YOLOv11 in crowd counting tasks. The process includes defining the datasets employed, evaluating each YOLO version on benchmark datasets, identifying appropriate evaluation metrics, configuring the experimental setup, and benchmarking against baseline models. This systematic approach ensures a holistic assessment of each YOLO model's accuracy, processing speed, and memory usage, collectively referred to as computational efficiency.

3.1 Evaluation and Comparison of YOLOv9, YOLOv10, and YOLOv11 Variants on the COCO Dataset:

For each YOLO version (YOLOv9, YOLOv10, and YOLOv11), their evaluation performance on the COCO dataset[14] will be provided. The corresponding tables show key metrics such as the Average Precision (AP), FLOPs, and model parameters, highlighting the trade-offs between model size, computational complexity, and detection accuracy. These performance metrics will guide the selection of the most appropriate variant for different crowd counting tasks, balancing between real-time efficiency and detection precision.

3.1.1 YOLOv9: As shown in table 1, YOLOv9 comes in multiple variants such as YOLOv9-T, YOLOv9-S, YOLOv9-M, YOLOv9-C, and YOLOv9-E, where each variant has a different number of parameters and FLOPs, affecting the AP at various IoU thresholds.

Table1. Performance Comparison of YOLOv9 Variants on the COCO Dataset[8]

Model	Test Size	AP_val	AP50_val	AP75_val	Param.	FLOPs
YOLOv9-T	640	38.3%	53.1%	41.3%	2.0M	7.7G
YOLOv9-S	640	46.8%	63.4%	50.7%	7.1M	26.4G
YOLOv9-M	640	51.4%	68.1%	56.1%	20.0M	76.3G
YOLOv9-C	640	53.0%	70.2%	57.8%	25.3M	102.1G
YOLOv9-E	640	55.6%	72.8%	60.6%	57.3M	189.0G

3.1.2 YOLOv10: The table 2 for YOLOv10 shows six variants ranging from YOLOv10-N to YOLOv10-X, with corresponding AP values, FLOPs, and inference latency for each model. This helps in selecting the best model for tasks requiring either faster inference or higher accuracy.

Table2. Performance Comparison of YOLOv10 Variants on the COCO Dataset[6]

Model	Test Size	#Params	FLOPs	APval	Latency
YOLOv10-N	640	2.3M	6.7G	38.5%	1.84ms
YOLOv10-S	640	7.2M	21.6G	46.3%	2.49ms
YOLOv10-M	640	15.4M	59.1G	51.1%	4.74ms
YOLOv10-B	640	19.1M	92.0G	52.5%	5.74ms
YOLOv10-L	640	24.4M	120.3G	53.2%	7.28ms
YOLOv10-X	640	29.5M	160.4G	54.4%	10.70ms

3.1.3 YOLOv11: The YOLOv11 table 3 includes variants from YOLOv11n to YOLOv11x, with details on mAP, speed (both CPU and GPU), number of parameters, and FLOPs. These metrics indicate the improvements in both speed and accuracy compared to previous versions.

Table3. Performance Comparison of YOLOv11 Variants on the COCO Dataset[9]

Model	size (pixels)	mAPval 50-95	Speed (ms) CPU ONNX	Speed (ms) T4 TensorRT10	params (M)	FLOPs (B)
YOLO11n	640	39.5	56.1 ± 0.8	1.5 ± 0.0	2.6	6.5
YOLO11s	640	47.0	90.0 ± 1.2	2.5 ± 0.0	9.4	21.5
YOLO11m	640	51.5	183.2 ± 2.0	4.7 ± 0.1	20.1	68.0
YOLO11l	640	53.4	238.6 ± 1.4	6.2 ± 0.1	25.3	86.9
YOLO11x	640	54.7	462.8 ± 6.7	11.3 ± 0.2	56.9	194.9

3.2 Datasets

The evaluation uses multiple benchmark datasets to ensure a diverse range of crowd environments, ranging from high-density urban crowds to sparse pedestrian areas. Each dataset was selected for its ability to simulate real-world crowd scenarios, such as occlusion and density variation. These datasets have been managed, preprocessed, and augmented using the Roboflow platform[30] to meet YOLO's specific requirements, including converting annotations to the YOLO format, resizing images, normalizing inputs, and applying data augmentation techniques. Notably, the majority of these datasets originally come in point annotation format, and significant effort was made to convert them into the

YOLO annotation format, which utilizes bounding boxes. This transformation was crucial to align the datasets with YOLO's requirements, involving considerable adjustments to localize individuals in the images accurately. The combined efforts of dataset management, preprocessing, and augmentation enhance the evaluation process by ensuring uniform, accurate data for detection and crowd counting across different scenarios.

3.2.1 SCUT-HEAD Dataset:

The SCUT-HEAD dataset is a large-scale head detection dataset collected and labeled for head detection tasks, particularly in indoor environments. It contains 4,405 images with 111,251 annotated heads, split into two parts:

- **Part A** consists of 2,000 images sampled from classroom monitor videos, with 67,321 heads annotated.
- **Part B** contains 2,405 images crawled from the internet, with 43,930 heads annotated. This dataset focuses on head detection in indoor settings, which makes it suitable for applications where face or body detection is not viable, such as when individuals have their backs to the camera [31]

3.2.2 ShanghaiTech Dataset:

The ShanghaiTech dataset is one of the largest datasets used for crowd counting, with 1,198 annotated images and 330,165 annotated heads. The dataset is divided into two parts:

- **Part A** includes 482 images that were randomly crawled from the internet and primarily contain highly congested scenes with large numbers of people.
- **Part B** consists of 716 images captured from busy streets in metropolitan areas of Shanghai, featuring relatively sparse crowd scenes. This dataset is widely used for evaluating crowd counting models due to its large scale and the varying densities and crowd compositions between Part A and Part B [19].

3.2.3 Mall Dataset:

The Mall dataset is a public dataset captured from a surveillance camera in a shopping mall. It contains 2,000 video frames at a resolution of 320×240, with diverse crowd densities ranging from sparse to crowded and a total of 62,325 annotated pedestrians. The dataset features challenging conditions such as lighting variations, reflections from glass surfaces, and severe perspective distortion, making it ideal for testing crowd counting models under real-world indoor settings [32].

3.3 Evaluation Metrics

To evaluate the YOLOv9, YOLOv10, and YOLOv11 models, we employ a range of metrics designed to measure accuracy, speed, and computational efficiency.

3.3.1 Detection Accuracy (mAP): Mean Average Precision (mAP) is the standard metric used to evaluate the accuracy of object detection models. For this study, mAP@50 (IoU threshold of 50%) will be used to measure how well each YOLO model detects and localizes individuals in the crowd. A higher mAP score indicates better detection accuracy.

3.3.2 Counting Accuracy (MAE and RMSE)

- Mean Absolute Error (MAE) calculates the average difference between predicted and actual crowd counts, providing a direct measure of counting accuracy.
- Root Mean Squared Error (RMSE) is a more sensitive metric that penalizes larger errors, giving insight into the consistency of the model's predictions across varying crowd densities.

3.3.3 Inference Speed (FPS): Frames per second (FPS) will be measured to determine how quickly each model processes images. This metric is crucial for real-time applications, where higher FPS indicates that the model is suitable for live crowd monitoring.

3.3.4 Computational Efficiency (GPU and Memory Usage): To assess the computational resource consumption of each model, GPU utilization and memory usage will be measured. This metric helps identify models that are more feasible for deployment in edge AI systems with limited hardware resources.

3.3.5 Robustness: The robustness of each YOLO model will be evaluated in various challenging conditions, including occlusion, low visibility, and varying crowd densities. The goal is to assess the models' ability to maintain high performance despite common real-world obstacles.

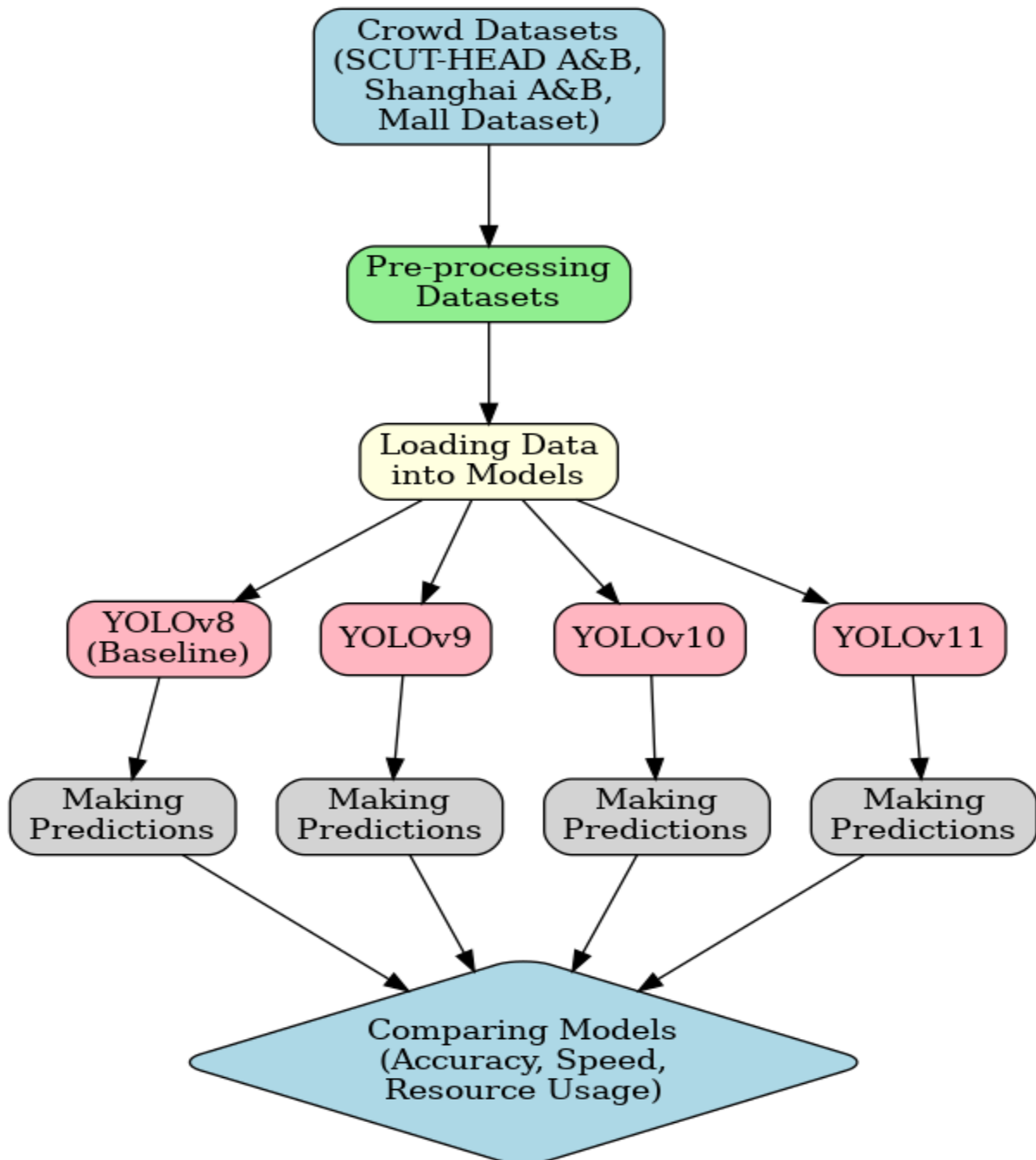


Figure 5: Proposed Methodology for YOLO-Based Crowd Counting

3.4 Experimental Setup

3.1 Hardware: All experiments will be conducted on Google Colab, utilizing its GPU resources to ensure fair comparisons between the models. This platform provides access to high-performance hardware, allowing for efficient training and testing of large-scale datasets while maintaining consistency across all experiments..

3.2 Software: The models will be implemented using PyTorch, which offers flexibility and efficiency for deep learning tasks. Pre-built libraries and transfer learning techniques will be leveraged to fine-tune YOLOv9, YOLOv10, and YOLOv11 on the datasets used in this study.

3.3 Baseline Models: YOLOv8 will serve as the baseline model for comparison against YOLOv9, YOLOv10, and YOLOv11. YOLOv8 is widely recognized for its balance of speed and accuracy, making it an appropriate reference point for assessing the improvements in newer YOLO versions.

4. Training and Testing Process

4.1 Baseline Testing: Before training, each YOLO variant (e.g., YOLOv9n, YOLOv10n, YOLOv11s) will be evaluated using pre-trained weights. This initial testing provides baseline performance data on metrics such as mAP, MAE, RMSE, and FPS across the datasets. These results will serve as a reference for assessing improvements after training.

4.2 Model Training: Each YOLO variant will be trained on the selected datasets for a fixed number of epochs (e.g., 10, 20, 50 epochs) to improve performance. Data augmentation techniques will be employed during training to enhance the model's ability to generalize across different environments. The best-performing weights will be saved for further evaluation.

4.3 Post-Training Testing: After training, the models will be re-evaluated on the same datasets. The post-training results will be compared with the baseline metrics to assess improvements in detection accuracy, crowd counting precision, speed, and computational efficiency. This phase provides insight into how well each YOLO model adapts to crowd counting tasks after fine-tuning.

4. Experimental Results:

Performance evaluation is ongoing due to the complexity and scale of the task. First, we are working with multiple datasets (ShanghaiTech A, ShanghaiTech B, SCUT-HEAD-A, SCUT-HEAD B, and Mall datasets), each with unique characteristics and challenges, requiring individualized testing for each model and dataset combination. Furthermore, we are evaluating several versions of YOLO models, starting from YOLOv8 (used as the baseline) to the most recent YOLOv11, which was released on 30 September 2024. Each version also includes multiple variants (n, s, m, l, and x), significantly increasing the number of evaluations needed to be completed.

Additionally, the evaluation process is computationally intensive, involving the measurement of key metrics such as detection accuracy (mAP), crowd counting precision (MAE, RMSE), inference speed (FPS), and computational efficiency (GPU and memory usage). This complexity is further exacerbated by the need to train, fine-tune, and test each model on every dataset, which is time-consuming.

Moreover, hardware limitations, particularly low computational power, further slow the evaluation process. Training deep learning models on large datasets demands significant processing power, and evaluation times are further prolonged on systems with limited hardware resources. Given that YOLOv11 has only recently been introduced, we are still completing the evaluations for this version, and the results will be communicated as soon as the testing is finalized. The comprehensive nature of this study, combined with the high volume of models and datasets and the hardware limitations, means that the evaluation process remains an ongoing effort.

5. Conclusion

This study provides a comprehensive performance evaluation of three key YOLO model variants—YOLOv9, YOLOv10, and YOLOv11—focusing on their applicability to crowd-counting tasks in diverse environments. The assessment was conducted using a variety of datasets, including SCUT_HEAD_partA, SCUT_HEAD_partB, ShanghaiTech Part A, ShanghaiTech Part B, and the Mall dataset, each characterized by varying levels of crowd density, occlusion, and environmental complexity. The models were evaluated in both their pre-trained (baseline) and post-training stages, using metrics such as mean average precision (mAP@50), mean absolute error (MAE), root mean squared error (RMSE), inference time, frames per second (FPS), and false positives/negatives to measure their performance.

References

- [1] A. Aldayri and W. Albattah, 'Taxonomy of Anomaly Detection Techniques in Crowd Scenes', *Sensors*, vol. 22, no. 16, Art. no. 16, Jan. 2022, doi: 10.3390/s22166080.
- [2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You Only Look Once: Unified, Real-Time Object Detection', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.
- [3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, 'Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation', in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81.
- [4] R. Girshick, 'Fast R-CNN', in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.
- [5] S. Ren, K. He, R. Girshick, and J. Sun, 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
- [6] A. Wang *et al.*, 'YOLOv10: Real-Time End-to-End Object Detection', May 23, 2024, *arXiv*: arXiv:2405.14458. Accessed: Oct. 20, 2024. [Online]. Available: <http://arxiv.org/abs/2405.14458>
- [7] 'Real Time Pedestrian Detection Using Modified YOLO V2', *IEEE Conf. Rec.*, 2020.
- [8] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, 'YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information', Feb. 29, 2024, *arXiv*: arXiv:2402.13616. Accessed: Oct. 20, 2024. [Online]. Available: <http://arxiv.org/abs/2402.13616>
- [9] Ultralytics, 'YOLO11  NEW'. Accessed: Oct. 20, 2024. [Online]. Available: <https://docs.ultralytics.com/models/yolo11>
- [10] 'Ultralytics YOLO11 Has Arrived! Redefine What's Possible in AI! by Abirami Vina'. Accessed: Oct. 20, 2024. [Online]. Available: <https://www.ultralytics.com/blog/ultralytics-yolo11-has-arrived-redefine-whats-possible-in-ai>
- [11] Ultralytics, 'Home'. Accessed: Oct. 20, 2024. [Online]. Available: <https://docs.ultralytics.com/>
- [12] 'Welcome to detectron2's documentation! — detectron2 0.6 documentation'. Accessed: Oct. 20, 2024. [Online]. Available: <https://detectron2.readthedocs.io/en/latest/>
- [13] M. Wadhwa, T. Choudhury, G. Raj, and J. C. Patni, 'Comparison of YOLOv8 and Detectron2 on Crowd Counting techniques', in *2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, Istanbul, Turkiye: IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/ISAS60782.2023.10391466.
- [14] T.-Y. Lin *et al.*, 'Microsoft COCO: Common Objects in Context', in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.
- [15] A. Neubeck and L. Van Gool, 'Efficient Non-Maximum Suppression', in *18th International Conference on Pattern Recognition (ICPR'06)*, Aug. 2006, pp. 850–855. doi: 10.1109/ICPR.2006.479.
- [16] A. Patwal, M. Diwakar, V. Tripathi, and P. Singh, 'Crowd counting analysis using deep learning: a critical review', *Procedia Comput. Sci.*, vol. 218, pp. 2448–2458, Jan. 2023, doi: 10.1016/j.procs.2023.01.220.
- [17] J. Redmon and A. Farhadi, 'YOLOv3: An Incremental Improvement', Apr. 08, 2018, *arXiv*: arXiv:1804.02767. Accessed: Oct. 20, 2024. [Online]. Available: <http://arxiv.org/abs/1804.02767>
- [18] N. Dalal and B. Triggs, 'Histograms of Oriented Gradients for Human Detection', in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA: IEEE, 2005, pp. 886–893. doi: 10.1109/CVPR.2005.177.

- [19] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, 'Single-Image Crowd Counting via Multi-Column Convolutional Neural Network', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 589–597. doi: 10.1109/CVPR.2016.70.
- [20] A. Menon, B. Omman, and A. S., 'Pedestrian Counting Using Yolo V3', in *2021 International Conference on Innovative Trends in Information Technology (ICITIIT)*, Kottayam, India: IEEE, Feb. 2021, pp. 1–9. doi: 10.1109/ICITIIT51526.2021.9399607.
- [21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, 'YOLOv4: Optimal Speed and Accuracy of Object Detection', Apr. 23, 2020, *arXiv: arXiv:2004.10934*. Accessed: Oct. 20, 2024. [Online]. Available: <http://arxiv.org/abs/2004.10934>
- [22] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, 'CSPNet: A New Backbone that can Enhance Learning Capability of CNN', in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580. doi: 10.1109/CVPRW50498.2020.00203.
- [23] C. Zhang, H. Li, X. Wang, and X. Yang, 'Cross-scene crowd counting via deep convolutional neural networks', in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 833–841. doi: 10.1109/CVPR.2015.7298684.
- [24] M. H. Kaka Khel *et al.*, 'Hybridized Yolov4 for Detecting and Counting People in Congested Crowds', in *2022 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Ottawa, ON, Canada: IEEE, May 2022, pp. 1–6. doi: 10.1109/I2MTC48687.2022.9806487.
- [25] M. Ahmadi, Z. Xu, X. Wang, L. Wang, M. Shao, and Y. Yu, 'Fast Multi Object Detection and Counting by YOLO V3', in *2021 China Automation Congress (CAC)*, Beijing, China: IEEE, Oct. 2021, pp. 7401–7404. doi: 10.1109/CAC53003.2021.9727949.
- [26] V. N. Rajeshirke and S. C. N., 'Human Detection in crowd using One-stage and Two-stage object detection models', in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India: IEEE, Jul. 2023, pp. 1–5. doi: 10.1109/ICCCNT56998.2023.10307765.
- [27] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, 'YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors', in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 7464–7475. doi: 10.1109/CVPR52729.2023.00721.
- [28] W. Liu *et al.*, 'SSD: Single Shot MultiBox Detector', in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.
- [29] J. C. Antony, Ch. L. S. Chowdary, N. Prabhu B, E. Murali, and A. Mayan, 'Advancing Crowd Management through Innovative Surveillance using YOLOv8 and ByteTrack', in *2024 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, Mar. 2024, pp. 1–6. doi: 10.1109/WiSPNET61464.2024.10533138.
- [30] 'Roboflow: Computer vision tools for developers and enterprises'. Accessed: Oct. 20, 2024. [Online]. Available: <https://roboflow.com/>
- [31] D. Peng, Z. Sun, Z. Chen, Z. Cai, L. Xie, and L. Jin, 'Detecting Heads using Feature Refine Net and Cascaded Multi-Scale Architecture', Nov. 24, 2018, *arXiv: arXiv:1803.09256*. doi: 10.48550/arXiv.1803.09256.
- [32] K. Chen, C. C. Loy, S. Gong, and T. Xiang, 'Feature Mining for Localised Crowd Counting', in *Proceedings of the British Machine Vision Conference 2012*, Surrey: British Machine Vision Association, 2012, p. 21.1-21.11. doi: 10.5244/C.26.21.

SESSION 6

Data Integration, Forecasting, and Recommender Systems



Real-Time Data Integration for Effective Congestion Forecasting

Messaoud Bechere¹✉, Abdelbasset Barkat²✉

¹ Laboratory of Informatics and its Application of M'sila (LIAM), University of M'sila, the Computer Science Department, M'sila, Algeria, messaoud.bechere@univ-msila.dz

² Laboratory of Informatics and its Application of M'sila (LIAM), University of M'sila, the Computer Science Department, M'sila, Algeria, abdelbasset.barkat@univ-msila.dz

Abstract

Congestion in smart cities is a challenging issue due to its impact on people's lives. Therefore, many researchers are developing solutions to this problem by taking advantage of new technologies, especially sensors and wireless network infrastructures. These technologies enable the collection of real-time information about many features directly or indirectly linked to traffic congestion.

In this paper, we propose a model for traffic congestion prediction based on real-time data, considering a set of relevant features such as traffic flow, incidents, holidays, and weather data. Our prediction classifies data into four categories: free flow, mild congestion, moderate congestion, and heavy congestion. After analyzing the results, the algorithm with the highest performance is XGBoost, achieving a prediction accuracy of 90%, followed by random forest and decision tree, both with a prediction accuracy of 89%.

Keywords: Traffic Congestion Prediction, Real-Time Data Integration, machine learning, Smart Cities

1. Introduction

Urban traffic congestion affects daily life, economic activity, and environmental sustainability worldwide for example, traffic jams cost the UK economy £6.9 billion in 2019, averaging 115 hours and £894 per motorist[1]. The total economic effect from 2013 is expected to reach £307 billion by 2030, including direct and indirect costs like wasted fuel and higher business fees[2]. This significant financial burden emphasizes the necessity for efficient traffic management solutions to decrease urban congestion and boost productivity.

Traffic control and forecasting models must be efficient to reduce congestion and improve urban mobility. These algorithms estimate traffic in real time, increasing road safety and transit efficiency. These technologies provide proactive traffic management and optimal route planning, allowing vehicles to avoid crowded routes and improve traffic flow and safety[3].

This research aims to create an all-encompassing traffic congestion prediction model by utilizing real-time data gathered from various sources in Westminster, UK. The dataset integrates traffic flow, incidents, holidays, and weather data to offer a comprehensive and real-time comprehension of traffic patterns. Machine learning algorithms improve traffic prediction by analyzing complex information and predicting congestion more accurately. Al-Selwi et al.[4] conclude that weather conditions substantially impact traffic flow and that their inclusion in prediction models leads to more robust and reliable traffic management strategies. Zhang et al.[5] found that traffic incident data and machine learning models considerably enhance peak-hour congestion prediction. These studies underline the need to combine several data sources for accurate traffic predictions. Taiwo et al.[6] showed the accuracy of several supervised machine learning techniques, including Classification Trees, Support Vector Machines, and Ensemble approaches, in forecasting traffic congestion. Their research validates that the integration of various machine-learning methodologies may significantly improve both the accuracy and efficacy of predictions. These studies underline the necessity of integrating several data sources and diverse computing methodologies to get accurate traffic forecasts.

Data for this study were collected from March to April 2024 within the defined geographic boundaries of Westminster. The area includes 70 junctions and 115 primary and secondary streets and is characterized by busy roads, historical landmarks, and commercial centers. Primary data sources include real-time traffic flow and incidents data from the TomTom API[7], road and intersection data from the OpenStreetMap platform[8], and weather data from the OpenWeather API[9].

Our approach involves collecting traffic flow data every 110 seconds for a high-frequency, comprehensive dataset. This method enables real-time traffic analysis and incident detection. The integration of weather factors such as temperature, rainfall, wind speed, and visibility is needed to see the impact of meteorological conditions on traffic flow.

Feature engineering, including the creation of time-based features and lagged speed variables, is critical for capturing temporal patterns and trends in traffic data. This study utilizes a range of machine learning methods, such as decision trees, random forests, and XGBoost, to determine the most efficient model for predicting traffic congestion. Tree-based algorithms have demonstrated significant potential because they can capture relationships among many attributes and effectively manage intricate data structures. Moreover, using categorical feature encoding and binarization algorithms guarantees that the model can efficiently handle and analyze various input types.

The objective of this study is to create a predictive model that can accurately forecast traffic congestion levels, by utilizing a diverse dataset from several sources in real-time and sophisticated machine learning algorithms, this work seeks to enhance the existing knowledge on traffic prediction and management.

This paper is organized as follows: Traffic congestion prediction models literature is reviewed in Section 2. Section 3 describes data collection, preparation. Section 4 discusses the model selection process and justification, Section 5 discusses the results obtained. Section 6 finishes the study and offers future research.

2. Literature review

Traffic congestion prediction has been studied using several methods and datasets. The study [4] used deep learning models, such as LSTM and GRU, to shed light on the role of weather data in traffic prediction. It also included weather parameters, such as temperature, humidity, and precipitation, into the models used to predict traffic flow. The results showed that the use of weather data greatly improves the accuracy of traffic flow predictions, particularly in adverse weather. However, other important factors, such as traffic incidents, were not taken into account.

The work [5] developed a machine-learning model for predicting urban traffic congestion, integrating incident data for better accuracy during peak hours. Decision trees and random forests were used to analyze traffic patterns and predict congestion levels. The integration of incident data considerably improved model performance, demonstrating the value of comprehensive datasets in traffic prediction models. The approach focused on peak hours and neglected off-peak hours which may limit the model's applicability.

In [6] analyzed the efficiency of three supervised machine learning algorithms Classification Tree, Support Vector Machine (SVM), and Ensemble (RUSBoosted) for traffic congestion prediction using a dataset from Kaggle. The study concluded that SVM achieved the highest prediction accuracy, followed by the Classification Tree and Ensemble methods. The study provided a comprehensive comparison of different machine learning algorithms, it lacked a detailed discussion on the influence of external factors such as weather and incidents on traffic congestion. The dataset used was relatively small, which may limit the generalizability of the findings; the study did not address the computational complexity of the models, which is an important consideration for real-time applications.

In the study [10] used a dataset from the US Traffic 2015 to present a machine-learning method for predicting vehicle traffic. In order to create prediction models, the study investigated a number of machine learning techniques, such as decision trees, support vector regression, and linear regression. This study successfully illustrated how several machine-learning algorithms may be used to traffic prediction. It did, however, mostly rely on historical traffic data, which would not have captured dynamism and abrupt shifts in travel patterns in real time. Moreover, weather and traffic incidents were not taken into account in the study, which is important for precise real-time forecasts.

Predicted road traffic congestion in [11] using the Random Forest method, which can handle huge datasets and multiple features. The research recommended merging traffic flow, incidents, and weather data to increase forecast accuracy. This study showed the Random Forest algorithm's traffic prediction powers. It didn't compare the method to other machine learning models, which may have helped explain its success.

The work [12] predicted traffic using sophisticated internet datasets, Random Forest, and LSTM algorithms. The study found that feature engineering, such as adding day and weekend components, greatly improved prediction models. The study proved feature engineering affects model performance. It focused on LSTM and Random Forest

algorithms without considering other potentially effective models. The study also did not examine real-time data integration, which is essential for traffic control systems.

Suggested employing LSTM networks to forecast traffic congestion, utilizing big data analytics to handle enormous amounts of traffic data in [13]. The study demonstrated that LSTM networks could capture traffic data temporal relationships, improving prediction accuracy. The study relied on historical traffic data rather than real-time updates, which are essential for congestion management. LSTM performance was not compared to other machine learning models.

[14] Compared decision trees, K-nearest neighbors (KNN), and random forests for traffic congestion prediction. The study found that random forests beat single classifiers in accuracy and resilience. The study thoroughly compared machine-learning methods. The study did not include weather and traffic accidents, which are essential for real-time congestion forecast.

The authors in [15] developed a traffic flow prediction model to estimate the probability of congestion in urban areas. The study used historical traffic data and machine-learning techniques to predict congestion levels accurately. It did not consider real-time data integration.

The paper [16] focused on predicting traffic congestion at urban intersections using machine learning algorithms. The study demonstrated that models incorporating traffic signal data and intersection-specific features significantly improved prediction accuracy. The study provided valuable insights into intersection-specific congestion prediction. It focused on static data from traffic signals and did not consider dynamic real-time data.

Introduced hybrid models that use machine learning techniques to predict short-term traffic congestion, the study [17] found that hybrid models might use algorithm strengths to increase prediction accuracy. The study did not examine real-time data integration, only short-term projections.

Using big data analytics to forecast traffic congestion, the study [18] integrated various data sources, like traffic flow, incidents, and weather, to develop a comprehensive forecasting model. The results indicated that big data analytics could significantly enhance real-time traffic management. The focus on big data may pose challenges related to data preprocessing and computational resources for practical applications.

The paper [19] demonstrated the efficacy of Long Short-Term Memory (LSTM) networks in predicting urban traffic congestion using IoT data, achieving accuracy levels between 84% and 95% based on road layouts. However, their focus was primarily on vehicle speed data, potentially missing other crucial traffic variables. Xue et al. [20] furthered this discourse by integrating environmental factors like weather into congestion prediction models using fuzzy neural networks, significantly improving predictive accuracy during adverse weather conditions.

3. Methodology

This research employs a multi-faceted approach to data collection, analysis, and model validation to predict traffic congestion. We collected these datasets from diverse sources, including traffic flow, incidents, and meteorological conditions. Following data collection, we applied preprocessing techniques to cleanse and structure the data, ensuring its readiness for in-depth analysis. We applied machine learning techniques utilizing decision trees, random forests, and XGBoost algorithms. These models were meticulously trained and tested to validate their effectiveness in forecasting traffic conditions

3.1. Data collection

The dataset was collected in Westminster, central London, inside the bounding box of 51.5111, -0.1549 southwest, and 51.5189, -0.1354 northeast Figure 1. It has bustling roadways, historical landmarks, and commercial centers. The target area has 70 junctions and 115 primary and secondary streets. Data from March and April 2024 involved real-time traffic flow and incident data from TomTom API, a Dutch GPS technology provider. OpenStreetMap platform provided road and junction data. Data was also obtained from the weather API. Weather factors like temperature, rainfall, wind speed, and visibility affect traffic.

Traffic flow data were collected every 110 seconds to capture dynamic traffic conditions. Additionally, traffic incident data were collected at the same time. Incidents or road closures produce fast traffic pattern changes, which are collected promptly and analyzed using a highly repeatable data collection method.

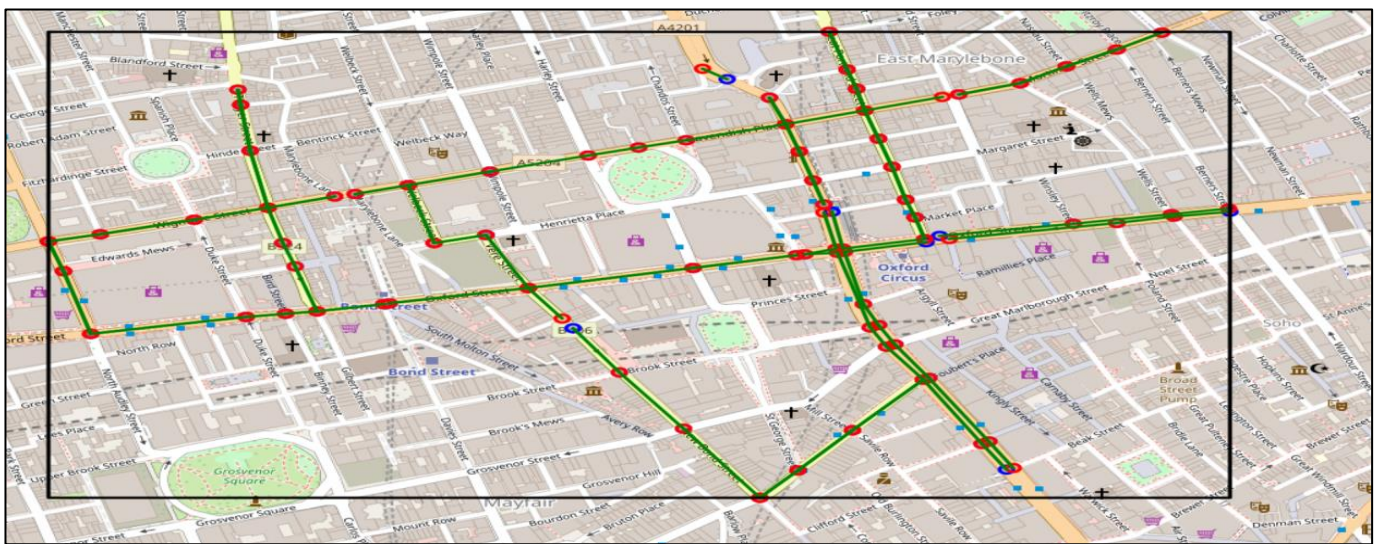


Figure 1. Geographic Scope and Data Collection Points in Westminster

About 3 million records were captured and stored on a PythonAnywhere[21] server. This study collected data from numerous sources at high frequency to create a comprehensive and detailed dataset. Traffic congestion research and forecasting are based on traffic flow, incidents, and meteorological data, enabling improved predictive models.

Table 1. The description of the dataset

Feature	Mean	Standard Deviation	Minimum	Maximum
Temperature	10.54 °C	3.19 °C	3.99 °C	20.86 °C
Visibility	9421.29 meters	1677.87 meters	800 meters	10000 meters
Wind Speed	18.75 km/h	8.19 km/h	2 km/h	48 km/h
Rain	0.1 mm	0.47 mm	0.0 mm	5.31 mm
Current Speed	13.12 km/h	4.55 km/h	1 km/h	28 km/h
Free Flow Speed	15.02 km/h	4.23 km/h	2 km/h	30 km/h
Current Travel Time	95.20 sec	65.02 sec	12 sec	2300 sec
Free Flow Travel Time	79.94 sec	50.82 sec	11 sec	508 sec

3.2. Evaluation of Traffic Congestion Levels

To assess the degree of congestion we used the equation (1) to calculate the Traffic Jam Level for each road segment by deriving the Speed Reduction Factor (SRF)

$$SRF = \frac{vf - vx}{vf} \quad (1)$$

Where:

SRF Is the Speed Reduction Factor

vf Is the Free Flow Speed

vx Is the Current Speed of traffic

Table 2 outlines the traffic jam levels determined based on different ranges of SR calculated for each segment:

Table 2. Classification of Traffic Jam Levels Based on SRF

Condition	Traffic Jam Level	Explanation
$SRF < 0.2$	Free Flow	The current speed is close to the free flow speed
$0.2 \leq SRF < 0.4$	Mild Congestion	Traffic is slightly slower than usual but still moving relatively smoothly.
$0.4 \leq SRF < 0.6$	Moderate Congestion	Traffic is noticeably slower with some delays expected.
$0.6 \leq SRF$	Heavy Congestion	Significant delays and slower traffic flow are observed.

3.3. Preprocessing

Table 1 shows the description of the dataset. Our data preprocessing included Data Transformation, Encoding Categorical Features, and Feature Binarization. Each step is critical for preparing predictive modeling data.

- **Data Transformation**

We started by converting different attributes to facilitate analysis. "date_Time" has been converted to DateTime format, allowing us to extract and create new time-related features. Specifically, we derive "DayOfWeek" and "IsWeekend". "DayOfWeek" indicates the day of the week on which we converted days from string to numeric (Sunday 0 to Saturday 6), and "IsWeekend" is a binary feature that determines whether a day is a weekend (1) or a weekday (0).

- **Encoding Categorical Features**

Encoding categorical features helps machine learning algorithms understand them. The 'IconCategory' was encoded with one-hot encoding for categorical data. This method converts category variables to binary, which machine learning systems can understand.

- **Feature Binarization**

We created the "PeakHour" feature to represent peak traffic hours. The IsAlternativeSegment feature indicates which segments represent alternative paths in certain situations.

For specific public holidays, our dataset improves on traffic congestion prediction using this binary feature.

3.4 Feature extraction

- **Traffic Jam Probability Estimation**

We calculated the probability of each traffic congestion level hourly for every road. Altering probability generated supplementary traffic congestion features. The four features with differing congestion levels were labelled 'prob_class_0', 'prob_class_1', 'prob_class_2', and 'prob_class_3'. The pivoted data frame was included into the original dataset to provide context for the items.

Algorithm: Traffic Jam Probability Estimation

Begin

Count occurrences of each traffic jam level for each combination

For each combination **do**

 | Count $\leftarrow$ number of occurrences for the combination

Calculate the total counts for each combination of StartPoint, EndPoint, Hour, Day, and Month

For each combination **do**

 | Total $\leftarrow$ sum of the combination

For each record **do**

 | Probability $\leftarrow$ Count / Total

Pivot data to get probabilities in separate features for each traffic jam level

End

- **Lagged Speed Feature**

To understand the change in speed over time, we created a new feature that contains the last speed in 110 seconds for each road. To achieve this, the data is first grouped by the combination of the startpoint and the endpoint. This ensures that the subsequent operations are performed within the context of each unique route rather than across the entire dataset. We isolate the speed values corresponding to each specific route by grouping the data.

The new feature we created is a powerful tool for time series analysis. It enables us to calculate speed differences, identify trends, and conduct further statistical analysis on how speed changes over time between routes.

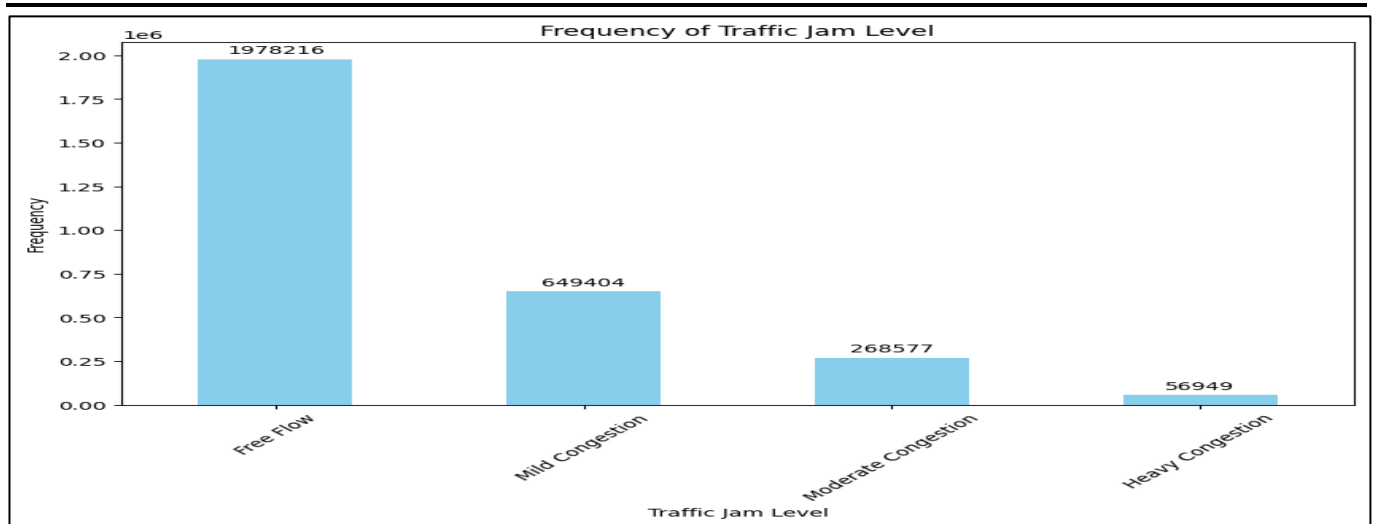


Figure 2. Data distribution

Figure 2 shows the data distribution by traffic jam level. The categories indicate traffic jam levels, from "Free Flow" to "Heavy Congestion." Nighttime circumstances considerably affect the overall distribution. While congestion is present, it is not the norm for all day. However, the less frequent congestion underscores the necessity for focused measures during peak hours and after incidents to preserve traffic flow.

4. Model selection

Time, holidays, and weather were used to develop an accurate road traffic jam forecasting model. This varied input dataset includes as many factors related to traffic flow dynamics as possible.

Features Extraction is essential for this research. Each of these features was selected based on their pertinence to traffic congestion forecasting. Temporal features include daily and weekly traffic cycles, whereas weather and traffic flow features elucidate the impact of external factors on driving behaviors. Features connected to incidents are essential for detecting congestion resulting from accidents or roadworks prevalent in urban traffic situations. Together, these features create a comprehensive model that can predict traffic congestion levels across different conditions and times.

We create a large dataset from the first module inputs. Using this dataset and machine-learning techniques, traffic patterns can be studied in detail. We have tested various algorithms, such as decision trees, random forests, and XGBoost. We applied several machine learning algorithms for better results. Despite our varied results, tree-based algorithms performed best. We found that tree-based algorithms perform better than others, and this returned for influenced by interactions between multiple factors; tree-based models automatically detect and leverage these interactions between features while building the trees. This capability makes them particularly capable of understanding complex patterns that other models might miss.

5. Results and discussions

The dataset consists of 20 features, which are free flow speed, free flow travel time, hour of day, day of week, peak hour, temperature, visibility, wind speed, amount of rain, use of alternative routes, holidays, and probability of jam in each hour, categorized by incident types (blocked road, road work, busy road due to rain, none). The dataset includes a single goal variable, traffic jam level, with four distinct values: free flow, mild congestion, moderate congestion, or heavy congestion.

The collection has 2,953,146 samples. We directly obtain the time and date properties from the system. We obtain weather data via the OpenWeatherMap API. OpenStreetMap provides roads data, including startpoint, finish point, length, type, and whether the road is one-way.

We use the TomTom API to obtain traffic data, which includes details like current speed, free flow speed, current travel time, free flow travel time, icon category, and delay.

Our analysis incorporates multiple machine learning models to predict different levels of traffic congestion ranging from 'Free' to 'Heavy.' Table 3 displays traffic forecasting model findings. The table compares Random Forest, XGBoost, and Decision Tree accuracy.

The accuracy rates of these models vary by approach and complexity. The most accurate is XGBoost at 90%. It handles diverse data types, complicated feature interactions, and overfitting, especially in imbalanced datasets, which contributes to its success.

Table 3. Accuracy of different models.

Classifier	Class	Precision	Recall	F1-Score	Accuracy
Random Forest	Free	0.94	0.96	0.95	89%
	Mild	0.79	0.76	0.78	
	Moderate	0.76	0.75	0.75	
	Heavy	0.86	0.83	0.85	
XGBoost	Free	0.94	0.96	0.95	90%
	Mild	0.80	0.76	0.78	
	Moderate	0.77	0.76	0.76	
	Heavy	0.87	0.85	0.86	
Decision Tree	Free	0.93	0.96	0.95	89%
	Mild	0.79	0.74	0.77	
	Moderate	0.76	0.72	0.74	
	Heavy	0.86	0.81	0.83	

As our dataset is imbalanced, therefore precision alone may favor the majority class. For a complete model performance assessment, we have drawn PR Curves for the analysis of performance. PR Curves of the output for the best scoring algorithms are shown in Figure 3, Class 0 means Free, Class 1 means Mild Congestion, Class 2 means Moderate Congestion, and Class 3 means Heavy Congestion.

Although the accuracy disparity between XGBoost (90%) and the Random Forest and Decision Tree models (both 89%) is minimal, it is crucial to evaluate additional performance metrics such as Average Precision (AP), which offers a more comprehensive assessment of model efficacy, particularly in imbalanced datasets. In this analysis, XGBoost exhibited superior performance in essential traffic congestion categories, attaining an AP of 0.99 for Free Flow and 0.94 for Heavy Congestion. These elevated AP values signify that XGBoost is especially adept at forecasting both the absence (Free Flow) and presence (Heavy Congestion). Considering the significance of these events for real-time traffic management, this robust performance in identifying critical congestion levels substantiates the preference for XGBoost over alternative models, notwithstanding the slight difference in overall accuracy.

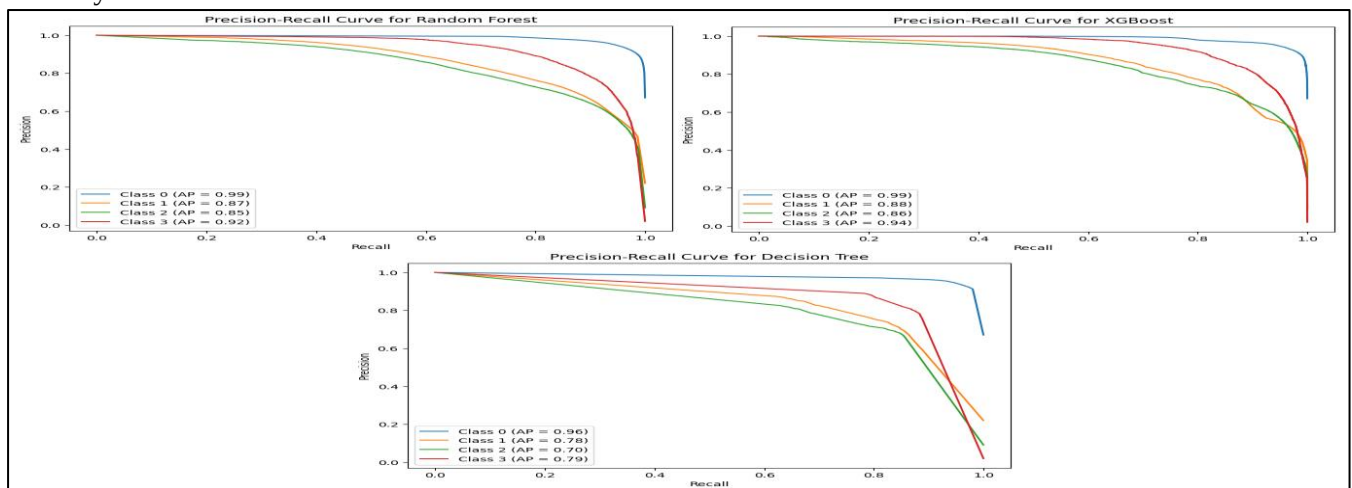


Figure 3. PR Curve

Figure 4 shows the confusion matrix for the best-performing model in our study, XGBoost. The diagonal elements of the matrix show each traffic congestion class's True Positive (TP) rates. The model successfully detected 96% of Class 0 (Free Flow) cases, suggesting good accuracy in recognizing smooth traffic. Class 1 (Mild Congestion) has a true positive rate of 76%, reflecting the model's reasonable accuracy in identifying slight congestion. Class 2 (Moderate Congestion) shows a true positive rate of 76%, demonstrating moderate performance in recognizing this level of traffic. The highest accuracy is observed for Class 3 (Heavy Congestion), with a true positive rate of 85%. The decision to present the confusion matrix for XGBoost highlights its superior performance across various metrics, making it the most reliable model.

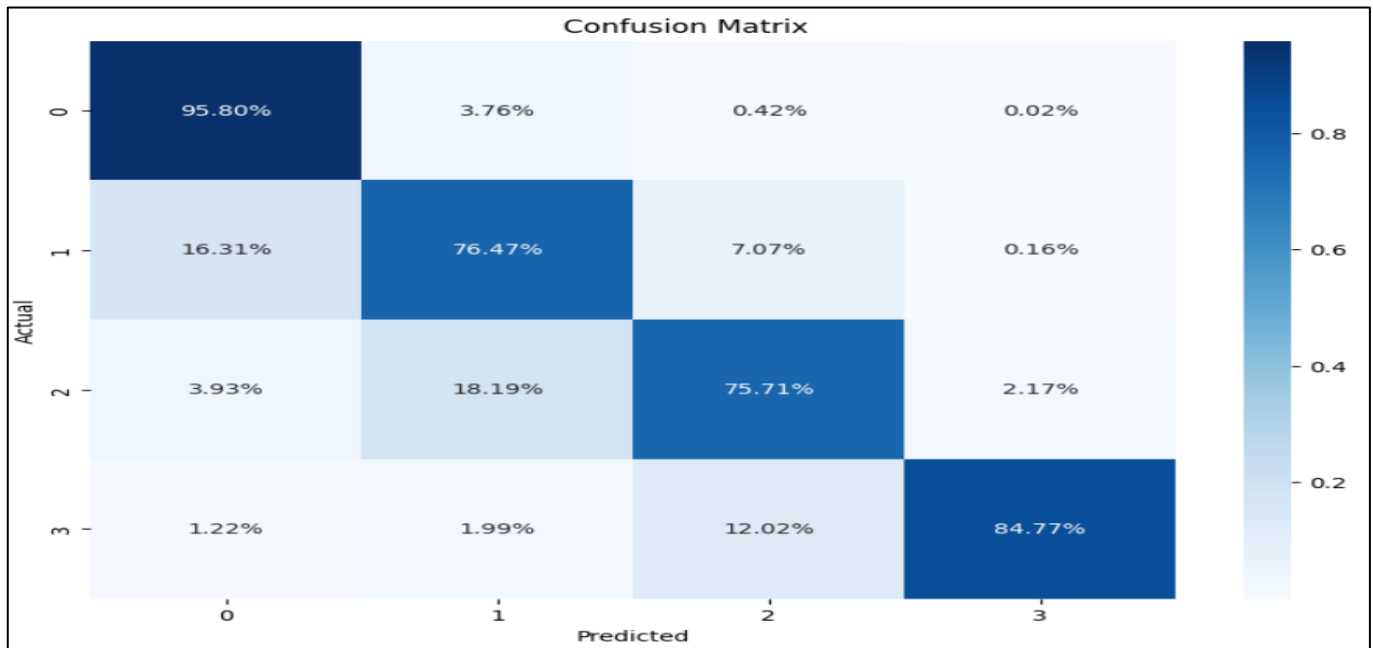


Figure 4. Confusion matrix for XGBoost

6. Conclusion

This study showed that incorporating traffic flow, incidents, holidays, and weather data improves urban traffic congestion forecast. We used XGBoost, Random Forest, and Decision Trees to create a predictive model that handles data complexity and has high prediction accuracy, as shown by XGBoost's 90% performance.

This study's results underscore the significance of including AP, particularly when assessing models on unbalanced datasets. Although the accuracy disparity among XGBoost, Random Forest, and Decision Tree is minimal, the significantly higher AP scores for XGBoost in critical congestion classes like Free Flow and Heavy Congestion indicate its better proficiency in forecasting key traffic events. This makes XGBoost the preferred choice for real-time traffic congestion forecasting, where predicting extreme traffic conditions with high precision is more critical than achieving marginal improvements in overall accuracy.

Our findings emphasize the importance of feature engineering. Time-based features and delayed speed variables have helped models accurately predict congestion levels by incorporating traffic flow temporal trends. Additionally, categorical feature encoding and binarization have greatly enhanced the models' capacity to process and interpret varied input sources.

Real-time traffic prediction is complicated by the computing needs of processing datasets and the requirement for models to react to quickly changing traffic circumstances, as shown by the research. However, the frequent data collecting and detailed analytic methodologies used in this study give a solid framework for addressing these problems.

This study offers a scalable solution to congestion and urban mobility, advancing traffic management. In future study we will integrate more real-time data, and use the models in diverse urban environments to evaluate and improve the prediction.

References

- [1] INRIX, “INRIX Global Traffic Scorecard: Congestion cost UK economy £6.9 billion in 2019,” Inrix. Accessed: Jul. 06, 2024. [Online]. Available: <https://inrix.com/press-releases/2019-traffic-scorecard-uk/>
- [2] INRIX, “Traffic Congestion to Cost the UK Economy More Than £300 Billion Over the Next 16 Years,” Inrix. Accessed: Jul. 06, 2024. [Online]. Available: <https://inrix.com/press-releases/traffic-congestion-to-cost-the-uk-economy-more-than-300-billion-over-the-next-16-years/>
- [3] P. Sun, N. Aljeri, and A. Boukerche, “Machine Learning-Based Models for Real-time Traffic Flow Prediction in Vehicular Networks,” *IEEE Network*, vol. 34, no. 3, pp. 178–185, May 2020, doi: 10.1109/MNET.011.1900338.
- [4] H. F. Al-Selwi, A. Bin Abd Aziz, F. S. Abas, N. A. Amir Hamzah, and A. B. Mahmud, “The impact of weather data on traffic flow prediction models,” *IJ-AI*, vol. 11, no. 4, p. 1223, Dec. 2022, doi: 10.11591/ijai.v11.i4.pp1223-1231.
- [5] K. Zhang, Z. Chu, J. Xing, H. Zhang, and Q. Cheng, “Urban traffic flow congestion prediction based on a data-driven model,” *Mathematics*, vol. 11, no. 19, p. 4075, 2023.
- [6] E. Taiwo, G. Ogunsanwo, and A. Olumuyiwa, “Traffic Congestion Prediction using Supervised Machine Learning Algorithms,” vol. 2, pp. 110–116, Jun. 2023.
- [7] “TomTom — Maps and Location Technology.” Accessed: Oct. 02, 2024. [Online]. Available: <https://www.tomtom.com/>
- [8] “OpenStreetMap,” OpenStreetMap. Accessed: Oct. 02, 2024. [Online]. Available: <https://www.openstreetmap.org/>
- [9] “Weather API - OpenWeatherMap.” Accessed: Oct. 02, 2024. [Online]. Available: <https://openweathermap.org/api>
- [10] A. Moses and R. Parvathi, “Vehicular Traffic analysis and prediction using Machine learning algorithms,” in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Feb. 2020, pp. 1–4. doi: 10.1109/ic-ETITE47903.2020.279.
- [11] H. Shenghua, N. Zhihua, and H. Jiaxin, “Road Traffic Congestion Prediction Based on Random Forest and DBSCAN Combined Model,” in *2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA)*, Zhangjiajie, China: IEEE, Jun. 2020, pp. 323–326. doi: 10.1109/ICSGEA51094.2020.00075.
- [12] T. M. Vishwasgowda and C. N. Chinnaswamy, “Traffic Prediction Using Machine Learning,” vol. 21, no. 3, pp. 161–165, 2023.
- [13] D.-H. Shin, K. Chung, and R. C. Park, “Prediction of Traffic Congestion Based on LSTM Through Correction of Missing Temporal and Spatial Data,” *IEEE Access*, vol. 8, pp. 150784–150796, 2020, doi: 10.1109/ACCESS.2020.3016469.
- [14] T. S. Tamir *et al.*, “Traffic Congestion Prediction using Decision Tree, Logistic Regression and Neural Networks,” *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 512–517, Jan. 2020, doi: 10.1016/j.ifacol.2021.04.138.
- [15] M. Pilamunga Poveda, J. M. Lavin, M. Aldás Sánchez, and D. Ríos Insua, “Traffic flow prediction model and probability of congestion,” *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 844, no. 1, p. 012008, May 2020, doi: 10.1088/1757-899X/844/1/012008.
- [16] T. Kelly and J. Gupta, “Predicting Traffic Congestion at Urban Intersections Using Data-Driven Modeling,” Jun. 07, 2024, *arXiv*: arXiv:2404.08838. Accessed: Jun. 26, 2024. [Online]. Available: <http://arxiv.org/abs/2404.08838>
- [17] M. Anjaneyulu and M. Kubendiran, “Short-Term Traffic Congestion Prediction Using Hybrid Deep Learning Technique,” *Sustainability*, vol. 15, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/su15010074.
- [18] F.-H. Tseng, J.-H. Hsueh, C.-W. Tseng, Y.-T. Yang, H.-C. Chao, and L.-D. Chou, “Congestion Prediction With Big Data for Real-Time Highway Traffic,” *IEEE Access*, vol. 6, pp. 57311–57323, 2018, doi: 10.1109/ACCESS.2018.2873569.
- [19] S. Majumdar, M. M. Subhani, B. Roullier, A. Anjum, and R. Zhu, “Congestion prediction for smart sustainable cities using IoT and machine learning approaches,” *Sustainable Cities and Society*, vol. 64, p. 102500, Jan. 2021, doi: 10.1016/j.scs.2020.102500.
- [20] Y. Xue, R. Feng, S. Cui, and B. Yu, “Traffic Status Evolution Trend Prediction Based on Congestion Propagation Effects under Rainy Weather,” *Journal of Advanced Transportation*, vol. 2020, no. 1, p. 8850123, 2020, doi: 10.1155/2020/8850123.
- [21] “Host, run, and code Python in the cloud: PythonAnywhere.” Accessed: Oct. 03, 2024. [Online]. Available: <https://www.pythonanywhere.com/>

A Concise Review of Online Benchmark Datasets for Recommender Systems

Amna Meddeb^{1*}, Lotfi Ben Romdhane²

¹ISITCom, MARS Research Lab LR17ES05, Tunisisa
amna.meddeb@isitc.u-sousse.tn

²ISITCom, MARS Research Lab LR17ES05, Tunisisa
lotfi.benromdhane@isitc.u-sousse.tn

ABSTRACT

Due to the vast expansion of online social network (OSN) data, strongly correlating with the growing OSN user base, data has become overwhelming and challenging to navigate. In response, researchers have tuned the forces to analyze and explore this data. Recommendation systems RS, represent a very known research domain dealing with this abundance of data, trying to provide users with personalized and relevant data, reducing their search efforts and frustrations when searching for information. However, the diverse range of recommendation models in the literature necessitates datasets that align with the recommendation models' requirements. Locating the suitable dataset remains challenging due to the sparsity and abundance of data within the network and the expensive time needed to analyse and extract the features of each data set. In this paper, we present an overview of online RS evaluation datasets, exploring their characteristics, features, description and references. Furthermore, to enhance accessibility, we provide the readers with two classifications of the datasets, providing a roadmap for swift identification of the most suitable dataset for their needs.

Keywords: Information systems, recommender systems, datasets, APIs

1. Introduction

With the immense growth of OSN data, exploring it has become overwhelming. Researchers have therefore collaborated to analyze this data and uncover its fascinating insights. One of the research domains dealing with this data is RS. It aims to provide users with relevant and personalized data helping them diminish the efforts and frustrations while searching for relevant data (information, content, location, etc).

In light of this, we mention that there is a wide variety of recommendation models that have been proposed and evaluated in the literature. Each model treats one or several aspects of RS, some models only consider raw information of users and items such as names, id, etc. Other models take advantage of social information. Consequently, the data used to evaluate the model needs to meet the model's requirements, for example, we cannot evaluate a model that considers users' neighbors' information with data that doesn't contain any social information. Thus, there is a need to allocate and find the appropriate data that can cover the recommendation model's requirements. Searching for the appropriate data is a hard task due to two factors. The first factor is related to the numerosity of available data that is sparse in the network making it hard to allocate, and the second issue is that these datasets need to be studied, to find the features they offer.

In this paper, we present an overview of online available RS evaluation datasets, going through their characteristics, descriptions, features, and main references. In additions, to make it obvious and easy

to search we present two classifications of the presented datasets, serving as a map to easily find the required dataset.

The rest of this paper is organized as follows. In section 2, we present an overview of online available datasets. In section 3 we present two classifications of the datasets. In section 4 we summarize this paper and point the interest over potential research directions.

2. Available evaluation data and APIs

This section presents an overview of benchmark datasets, other available datasets, and APIs.

Benchmark datasets and available datasets are presented in table 1 (Accessible data links can be found in the footnotes.) Benchmark datasets are commonly used within the research community to evaluate and compare the performance of different models and algorithms on a specific task. The other available datasets are datasets that are potential future benchmark datasets.

Table 1: Online available Datasets

Dataset	Ref	Details	Description	Features
MovieLens 25M ¹	[1]	25M ratings 1 million tags 62K movies 162K users.	MovieLens is a website equipped with a RS that asks users to provide their feedback in the form of ratings in order to receive personalized movie recommendations. MovieLens dataset contains users' preferences for movies expressed via a tuple composed of ratings from 1 to 5 for a rated movie.	Review: Movie id / Movie title / Movie genre / User-id / Ratings [0-5]
MovieLens 20M	[2], [3]	20M ratings 465,000 tag 27,000 movies 138,000 users.		
MovieLens 10M	[4]	10 M ratings 100,000 tag 10,000 movies 72,000 users.		
MovieLens 1M	[5]	1 M ratings 6000 users 4000 movies		
MovieLens 100K	[5]	100,000 ratings 1000 users 1700 movies		
Heterec-movielens-ik-v2 ²	[6]	2113 users 10197 movies 20 movie genres 4060 directors 95321 actors 72 countries 13222 tags 855598 ratings	Heterec-ML represent an enriched version of movielens, where enriching data comes from Internet Movie Database (IMDb) and Rotten Tomatoes movie review systems	users' informations : ratings/ tags movie information : genre/ directors/ actors/ contries/ location/ tags/ ratings
Movie tweet-ing ³	[7]	921,398 ratings 71,707 users 38,018 items	MovieTweatings contains users' reviews of movies that are recorded via users' well-structured tweets.	user: id/ twitter id items: movie id/ movie- title/ movie year/ genre/sub-genre ratings: user-id/ movie-id/ rating/ rating-time

¹<https://grouplens.org/datasets/movielens/>

NB: this is the same link for databases 2, 3, 4, and 5

²<https://grouplens.org/datasets/hetrec-2011/>

³<https://github.com/sidooms/MovieTweatings>

Netflix Prize data ⁴	[2]	480189 users. 17770 movies	These datasets collect information from Netflix, which is	Users ratings (CustomerID, Rating, Date) Movie: ID/ YearOfRelease/ Title
Netflix Movies and TV Shows ⁵	[6]	8807 movies and tv shows 70% movies and 30 % tv shows	a website that allows users to watch films and television shows from various genres online.	Movies or tv shows : ID/ type (movie or tv show)/ title/ director/ actors/ country/ adding to netflix date/ year of release/ rating / duration or number of seasons
Goodreads ⁶	[3]	1,561,465 items 808,749 users 225,394,930 interactions	Collected from Goodreads book review website and it contains reviews, items' attributes and different types of users' interaction, from adding to the "shelf" to rating to reading.	Review : user_id / book id/ review_id/ isRead/ rating/ date added/ date_updated/ read at/ started_at book attributes: title, isbn graph of similar books
Book-crossing (BX)	[6], [5], [8]	278,858 users 1,149,780 ratings 271,379 books.	Collected in a 4-week crawl in 2004 from the BookCrossing community with permission.	user: ID/ Location/ Age books: ISBN/ title/ author/ year of publication/ publisher/ cover_img Ratings: user id/ book_id/ 0..10
LibraryThing ⁷	[9]	73,882 users 337,561 items 979,053 ratings/feedback 120,536 social relations	This dataset include ratings as well as social relationships between users. The dataset combine data from LibraryThing (a book review website) and Epinions.	Review: product name/ product-category/ stars/ helpfulness/ time/ comment items : price users and social relation between users
Douban (sample) ⁸	[10]	785 users 5520 music 5623 book 170 film 830 book reviews 1277 music review 1073 film review	Douban is a Chinese online database and social network that allows users to watch and create content related to films, books, music, events, and activities in Chinese cities.	Reviews: user_id/ item_id/ rating/ label/ comment/ time/ id/ UID Book: ids/ UID Movies: id/ name/ director/ summary/ writer/ country/ pubdate/ language/ imdb/ numraters/ rating/ tag/ full name, categoryID Music: id User: id/ living place/ joint_time/ self_statement
Amazon Product Review ⁹	[3]	82.83M ratings 20.98M users 9.35M items	This data set contains users' reviews of Amazon products, there reviews are composed of users' ratings in a scale from 1 to 5. This dataset can be used as a hole or just a part of it since items are grouped per category and each category can form a sub-dataset. (all categories are available on the provided link).	Review: reviewerID/ asin (review id)/ reviewer name/ helpful/ review text/ overall (ratings) 0..5 /summary/ unix-Time / review time items : product image (and CNN features)/ price/ category/ salesRank. item-to-item relationships (e.g. "people who bought X also bought Y")

⁴<https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data>

⁵<https://www.kaggle.com/datasets/shivamb/netflix-shows>

⁶<https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home>

⁷https://cseweb.ucsd.edu/~jmcauley/datasets.html#social_data

⁸<https://www.kaggle.com/datasets/fengzhujoe/douban-datasetratingreviewside-information>

⁹https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

Amazon Question and Answer Data ¹⁰	[8]	1.48M questions 4,019,744 answers 309,419 Labeled yes/no questions 191,185 unique products with questions	This dataset contain questions and answers about products from the Amazon dataset. This data set is mostly used for message recommendation more precisely for chatbot training.	Q/A : Id/ questionType : Yes or NO/ answerType/ answerTime/ unix-Time/ question/ answer
ModCloth ¹¹	[11]	Reviews: 99,893 Items:1,020 Users: 44,783	These datasets are collected in the context of bias study [11], they contain attributes about products sold on ModCloth and Amazon which may be sources of bias in recommendations.	items : item-id/ model attr/ category/ brand/ product images/ item sizes user : user-id/ user identity/ gender/user_attr review : rating/ timestamp/ year/ split
Amazon Marketing ¹²	[11]	1,292,954 reviews 9,560 items 1,157,633 users		items : item-id/ model attr/ category/ brand/ product images/ item sizes user : user-id/ user identity/ gender/user_attr review : rating/ timestamp/ year/ split
Epinion ¹³	[12]	116,260 users 41,269 items 181,394 ratings/feedback 181,304 social relations	The Epinions is a trust network dataset where each user have a profile, ratings, and trust relations. A rating comprises 5 elements the product name and category, the rating score, the time when the rating is created, and the helpfulness of this rating.	Review : product name/ product-category/ stars/ helpfulness/ time/ comment items : price users and social relation between users
Google Local Reviews (2021) ¹⁴	[13]	Reviews: 666,324,103 Users: 113,643,107 Businesses: 4,963,111	The dataset is extracted from google maps. It contains information about businesses and users' reviews of these businesses.	Review : user_id/ name/ rating / time/ text/ pics/ resp/ gmap_id Business : name / address/ gmap_id/ description/ latitude/ longitude/ category/ avg_rating /num of_reviews/ price/ open hours/ MISC information/ current status of the business (e.g., permanently closed)/ relative_results - relative businesses recommended by Google/ url of the business
Yelp ¹⁵	[14]	6,990,280 reviews 150,346 businesses 200,100 pictures 11 metropolitan areas	Yelp is an application available on web and mobile that publishes crowd-sourced reviews about businesses. It also operates as a table reservation service.	Available on request on the Yelp Website

¹⁰<https://cseweb.ucsd.edu/jmcauley/datasets/amazon/qa/>

¹¹<https://github.com/MengtingWan/marketBias>

¹²<https://github.com/MengtingWan/marketBias>

¹³https://cseweb.ucsd.edu/jmcauley/datasets.html#social_data

¹⁴<https://jiachengli1995.github.io/google/index.html#code>

¹⁵<https://www.yelp.com/dataset>

Food.com Recipe Review Data ¹⁶	[15]	231,637 recipes 226,570 users 1,132,367 reviews	These datasets contain recipe details and reviews from Food.com	Ratings: user_id/ recipe_id/ date/rating/ Recipe: id/Name/ Description/ Ingredients/ Directions/ cooking techniques/ category (tags)/ Nutrition information users: id/ cooking techniques/ recipes interacted with/ number of reviewed recipes/ ratings per recipes/ total number of ratings
(Douban Conversation Corpus) ¹⁷	[16]	session response pairs : Training set: 1M Test set: 10k	Douban Conversation Corpus, contain a training data set, a development set and a test set usually used to train a chatbot.	Label/ conversation utterances/ response
(Sina) Weibo-COV 1.0 ¹⁸	[17]	20 million Weibo active user pool 65,175,112 tweets filtered from 2,615,185,101 original tweets by keywords	Sina Weibo is a Chinese microblogging website. Launched on 14 August 2009, and is one of the biggest social media platforms in China.	Posts : _id,user_id / crawl_time/ created_at/ like_num/ repost_num/ comment_num/ content/ origin_weibo/ geo_info Users features: user_id/ gender/ province/ city/ birthday/ fans_num/ vip_level/ crawl_time
Cross network data Youtube and Twitter	[18]	38,540 provide YouTube account 39,400 provide Twitter account 11,850 provide both accounts	The authors crawled a new dataset with user account linkage between YouTube and Twitter. (with the help of Google+)	Youtube: Upload history/favorite videos/ playlists/ videos information Twitter: followees, profiles
ZhihuRec-100M ¹⁹	[19]	100M interactions collected within 10 days, 798K users, 165K questions, 554K answers, 240K authors, 70K topics, more than 501K user query logs.	ZhihuRec dataset is collected from a knowledge-sharing platform (Zhihu), Zhihu is a forum website where questions are created, answered, edited, and organized by its users.	users' clicks and impressions users' queries users' features: id/ register timestamp/ gender/ login frequency/ followed topics/ followed questions/ answers/ questions/ comments/received thanks/ received comments/ received likes/ received dislikes/register type/ register platform/ used device information/ platform/ province/ city/ followed topic ids
ZhihuRec-20M	[19]	20M user-answer impression logs		answers features: id, question id/ anonymous or not/ author ID / labeled high-value answer or not/ recommended by the editor or not/ create timestamp/ media/ thanks/ likes/ comments/ collections/ dislikes/ reports/ helpless/ token ids in the answer/ topics ids of the answer

¹⁶<https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions>

¹⁷<https://www.dropbox.com/s/90t0qtj9ow20ca/DoubanConversationCorpus.zip?dl=0>

¹⁸<https://github.com/nghuyong/weibo-cov>

¹⁹<https://github.com/THUIR/ZhihuRec-Dataset>

ZhihuRec-1M	[19]	1M user-answer impression logs	questions features: id/ time/ answers/ followers/ invitations/ comments/ token ids in the questions/ topic ids authors features: id/ is excellent or not/ followers/ is excellent answerer or not ids of topics Tokens: token id and associated word vectors (obtained by training a word2vec model)
-------------	------	--------------------------------	---

3. Classification

In this section we introduce two classifications based on two criteria: the type of the recommended object 2 and the dataset features 3. These classifications will play the role of a map, helping researchers easily find their targeted needs in online available datasets.

Table 2: Classification based on the type of the recommended object.

Recommended object	Databases
Movies	MovieLens databases, Heterec-Movielens-ik-v2, Movie tweeting, Netflix prize data, Netflix Movies and TV shows, Douban (sample)
Books	Goodreads, BX, Library thing, Douban (sample),
Music	Douban (music)
Products	Amazon product Review, Amazon Question and Answer, ModCloth, Amazon marketing, Epinion
Location	Google Local Reviews, Yelp
Recipes	Food.com Recipe Review data
NLP	Douban conversation corpus, ZhihuRec datasets
Content	(Sina) weibo-cov1.0, Cross network data

The second classification 3 help researchers find the database they need focusing on defining their required features and searching for the database that contains those features. In Table 3 we distinguish 4 main classes of features. The first class of features is reviews' features where a review represents users' feedback and can be in a simple form or enriched with supplemental information such as review time, place, etc. The second class is related to items' features, the third concerns users' features, and the fourth class represents social features, and it englobes items and users' social features, that can be found in items' and users' networks respectively.

Table 3: Classification based on the databases' features

	Reviews features		Items features	Users features	Social features	
	Raw (ratings)	Enriched reviews			Items network	Users network
MovieLens datasets	X		X			
Heterec-Movielens-ik-v2	X		X	X		
Movie tweeting		X	X	X		
Netflix prize data		X	X			
Netflix Movies and TV shows	X		X			
Goodreads		X	X		X	
BX	X		X	X		
LibraryThing		X	X			
Douban (sample)		X	X	X		
Amazon product Review		X	X		X	
ModCloth		X	X	X		

Amazon marketing		X	X	X		
Epinion		X	X			X
Google Local Reviews		X	X			
Yelp						
Food.com Recipe Review data		X	X	X		
(Sina) weibo-cov1.0		X (implicit)	X	X		
Cross network data		X (implicit)	X	X		X
ZhihuRec datasets		X (implicit)		X		

In table 3 we notice that, while almost all datasets contain review and basic information of users and items, a limited number of datasets provide social features. Items' social features can be extracted implicitly through item tags, users' social networks, etc. Users' social information is generally extracted through their online social networks' profiles, accessing such information is very challenging due to the non-availability of most OSN data and that is mainly explained by their legacies that take the responsibility to protect users' private data.

4. Conclusion

In this paper, we tackle the challenge of efficiently selecting a suitable dataset, to save researchers valuable time. We have provided an overview of datasets available and accessible online for the evaluation of recommendation models. Furthermore, we have categorized the datasets into two classifications based on features and recommended object types. These classifications serve as a navigational guide, enhancing the ease and clarity of the dataset search process. In light of helping researchers find the adequate dataset, we find it interesting to further analyse the inner characteristics of databases, including sparsity, distribution, bias, etc. and a comparison between databases based on the obtained results will be interesting, allowing researchers to allocate the most appropriate database while addressing a specific issue, like sparsity or bias.

References

- [1] V. Radhika and K. Swaraj, "Movie genre prediction and recommendation using deep visual features from movie trailers," in *2020 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pp. 1–6, IEEE, 2020.
- [2] I. Shenbin, A. Alekseev, E. Tutubalina, V. Malykh, and S. I. Nikolenko, "Recvae: A new variational autoencoder for top-n recommendations with implicit feedback," in *Proceedings of the 13th international conference on web search and data mining*, pp. 528–536, 2020.
- [3] M. Stratigi, E. Pitoura, and K. Stefanidis, "Squirrel: A framework for sequential group recommendations through reinforcement learning," *Information Systems*, vol. 112, p. 102128, 2023.
- [4] X. Jia, Y. Dong, F. Zhu, Y. Xin, and J. Qian, "Preference-corrected multimodal graph convolutional recommendation network," *Applied Intelligence*, vol. 53, no. 4, pp. 3947–3962, 2023.
- [5] K. Rama, P. Kumar, and B. Bhasker, "Deep autoencoders for feature learning with embeddings for recommendations: a novel recommender system solution," *Neural Computing And Applications*, vol. 33, pp. 14167–14177, 2021.
- [6] H. Zhang, Y. Sun, M. Zhao, T. W. Chow, and Q. J. Wu, "Bridging user interest to item content for recommender systems: an optimization model," *IEEE transactions on cybernetics*, vol. 50, no. 10, pp. 4268–4280, 2019.

- [7] S. Kumar, K. De, and P. P. Roy, "Movie recommendation system using sentiment analysis from microblogging data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 915–923, 2020.
- [8] C. Park, D. Kim, X. Xie, and H. Yu, "Collaborative translational metric learning," in *2018 IEEE international conference on data mining (ICDM)*, pp. 367–376, IEEE, 2018.
- [9] T. Zhao, J. McAuley, and I. King, "Improving latent factor models via personalized feature projection for one class recommendation," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 821–830, 2015.
- [10] F. Zhu, Y. Wang, C. Chen, G. Liu, and X. Zheng, "A graphical and attentional framework for dual-target cross-domain recommendation.," in *IJCAI*, pp. 3001–3008, 2020.
- [11] M. Wan, J. Ni, R. Misra, and J. McAuley, "Addressing marketing bias in product recommendations," in *Proceedings of the 13th international conference on web search and data mining*, pp. 618–626, 2020.
- [12] D. Mumin, L.-L. Shi, L. Liu, and J. Panneerselvam, "Data-driven diffusion recommendation in online social networks for the internet of people," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 1, pp. 166–178, 2020.
- [13] A. Yan, Z. He, J. Li, T. Zhang, and J. McAuley, "Personalized showcases: Generating multi-modal explanations for recommendations," *arXiv preprint arXiv:2207.00422*, 2022.
- [14] X. Wang, X. Yang, L. Guo, Y. Han, F. Liu, and B. Gao, "Exploiting social review-enhanced convolutional matrix factorization for social recommendation," *IEEE Access*, vol. 7, pp. 82826–82837, 2019.
- [15] B. P. Majumder, S. Li, J. Ni, and J. McAuley, "Generating personalized recipes from historical user preferences," *arXiv preprint arXiv:1909.00105*, 2019.
- [16] L. Li, C. Li, and D. Ji, "Deep context modeling for multi-turn response selection in dialogue systems," *Information Processing & Management*, vol. 58, no. 1, p. 102415, 2021.
- [17] W. Yang, Z. Wu, N. Y. Mok, and X. Ma, "How to save lives with microblogs? lessons from the usage of weibo for requests for medical assistance during covid-19," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2022.
- [18] L. Sang, M. Xu, S. Qian, M. Martin, P. Li, and X. Wu, "Context-dependent propagating-based video recommendation in multimodal heterogeneous information networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 2019–2032, 2020.
- [19] H. Xue and D. Zhang, "A recommendation model based on content and social network," in *2019 IEEE 8th Joint international information technology and artificial intelligence conference (ITAIC)*, pp. 477–481, IEEE, 2019.

Trust in Normative Multi-agent Systems: Survey

Guerfi Imane^{1✉}, Toufik Marir², Maroua Bouzid³

¹ Research Laboratory on Computer Science's Complex Systems (ReLa(CS))², University of Oum El Bouaghi, Oum El Bouaghi, Algeria

imaneguerfiogx@gmail.com

² Research Laboratory on Computer Science's Complex Systems (ReLa(CS))², University of Oum El Bouaghi, Oum El Bouaghi, Algeria

marir.toufik@univ-oeb.dz

³ University of Caen Normandie, GREYC Laboratory, 14000 Caen, France
maroua.bouzid-mouaddib@unicaen.fr

Abstract

Multi-agent systems are widely used to develop complex systems. Such systems can be open, and the agent which composed them can be heterogeneous. Hence, a serious problem of trust can appear. Furthermore, normative multi-agent systems are a recent field of research. It uses the norms to control the agents' behaviours. Obviously, the norms can contribute positively to trusting inert agents because they control their behaviours. However, it is also possible to use the norms as a weak point in this category of multi-agent systems to influence trust. This paper presents a survey about trust in normative multi-agent systems to understand the main knowledge gaps in the field.

Keywords: Multi-Agent Systems, Normative Multi-Agent Systems, Trust.

1. Introduction

Recently, research into normative multi-agent systems (NorMAS) has been interesting. Knowing that these systems are the result of combining multi-agent systems and normative ones [1]. They allow controlling agents' behaviours and interactions using common norms and rules [2]. Consequently, deciding which actions are permitted or prohibited will be possible.

A multi-agent system (MAS) is a set of agents that interact with each other to solve a complex problem. An agent is an autonomous entity located in a specific environment, and it captures data from this later to decide based on the goal [3]. Also, autonomous agents are characterized by several attributes that allow them to achieve goals and make decisions, like reactivity, proactivity, and social ability [4].

A normative system is a system controlled by a set of norms. Norms can be laws, social rules, or predefined procedures that describe the obligatory, prohibited, and permitted actions [2] [5]. In computer science, normative systems are defined as systems in which norms play a role in their behaviour and require normative concepts to be described or specified [6]. Consequently, in the multi-agent context, the norms control the agents' behaviours and interactions. It allows us to talk about normative multi-agent systems.

An agent, as an autonomous entity, can have goals that are incompatible with the other agents' goals. In some extreme situations (like in open multi-agent systems), the agents can join or leave the system during runtime. As a consequence, agents' cooperation abilities are limited by knowing the goals of other agents in the system. Hence, trust must exist between the agents to make the interaction more useful [7] [8]. In [9], trust is considered an important element in both human and agent interactions.

This paper aims to provide a synthesis of trust in normative multi-agent systems and to analyze knowledge gaps in the field, identifying future research directions. The remainder of this paper is organised as follows: The next section, we target the study of trust in multi-agent systems because it provides a foundational understanding necessary for contextualizing trust within normative multi-agent systems.. In section 3, we present the related works that targeted trust in normative multi-agent systems before presenting a conclusion in section 4.

2. Trust in Multi-Agent Systems

Many researchers have tried to clarify and define the concept of trust, but there has yet to be a consensus [10]. Because it was studied in many disciplines, and each discipline has its own specific definition [11] [12]. Despite the variety of definitions, some characteristics are common across most of them. In fact, there is a strong relationship between trust and reliability or performance on the one hand [13]. On the other hand, trust is a cognitive process based on beliefs and expectations, influenced by the history of encounters. It is a subjective expectation about another's future behavior, influenced by previous experiences [14]. In [15] Relational trust is a belief-based cognitive process where a relational party commits to achieving goals despite unknown results and risks. Trust is also influenced by reputation, a perception created through past actions about an entity's intentions and norms. Trust depends on reputation, based on personal experiences and others' experiences [12]. On the other hand, [16] considers trust the basis for building reputation. In fact, direct personal relationships are built, and the level of dependability and reliability of trust is determined. Then, reputation is built as a classifying mechanism.

Audun Jøsang et al., [17] suggest that trust and reputation are interconnected and integrated, with trust influenced by personal experiences and interactions. In the absence of personal experience, reputation is used to evaluate an individual's standing in society. Trust is also a fundamental concept in security, as it serves as the foundation of security in multi-agent systems. Furthermore, security techniques can be used to reinforce trust.

In a multi-agent system, agents interact with each other to achieve their goals [8]; therefore, each agent has a goal they want to achieve by relying on another agent. This interaction needs to be motivated to achieve the agents' goals. In other words, the interaction between agents requires the existence of trust [18], making trust a major issue in MAS [19]. Trust can be understood by answering the question, "Why does an agent trust another agent? ". In [20], it is stated that trust in a MAS is "the truster (*i*) trusts the trustee (*j*) that the precondition *a* (*p*) holds and to bring about (*c*). " Indeed, the agent depends on another agent because the latter will achieve their goal under certain conditions.

Similarly, in [11], trust is defined as "the truster (*i*) trusts the trustee (*j*) to perform the content (*c*). " Thus, trust in MAS operates at the individual level, that is, at the level of agents. Agents are mental entities that use subjective measures of the principle of utility to decide on a potential partner for future interactions [21].

On the other hand, [22] considers that trust in a multi-agent system operates at both the individual and system levels, complementing the other to enhance trust in MAS. Trust at the individual level reflects trust in a particular individual, built through evaluating their past behaviours and interactions. This is done using direct trust and individual reputation to determine future interactions and social and cognitive methods to assess an agent's motivation and desire. Conversely, at the system level, trust is maintained through mechanisms that impose trust and

reputation across the system, alongside security measures that verify agent identities and ensure the integrity of interactions. These mechanisms serve as information sources for trust [7], and the agent's role in the system is also a source of information for trust [21].

Information sources of trust are widely used in trust models to provide agents with data to assist them in selecting an interaction partner. As described in [21], a coordination tool is proposed for agents in an open multi-agent system to make decisions about partners by classifying and assigning roles based on trust after monitoring agent interactions. This provides an impression or expectation of their behaviour and ability to perform certain tasks.

In [23], they integrated a multi-agent system with blockchain technology using smart contracts, where agents' tasks are stored for execution. Also, in [24], a model based on MAS and smart contracts was proposed to facilitate information exchange between agents. Communication between agents in the Internet of Vehicles (the vehicle agent and the service provider agent) was established using smart contracts to record data and outline task agreements and conditions, this facilitates partner selection for future interactions based on past performance history and tasks by the agent and achieves communication between agents. Meanwhile, in [25], they suggested a trust framework for decision-making that depends on measuring trust by accounting for agents' short and long-term behavioural histories, providing insights into their behaviours and trust levels.

Moreover, in [20], a framework called "Trust Computational Logic" was proposed to measure potential trust partners in MAS. This framework uses a "logical language" to define semantic trust structures (truster, trustee, behaviour, goals, and commitments), system descriptions, and trust definitions, and measures trust using vector-based interpreted systems with a behaviour library to compare observations and perceptions in a general matrix form.

In [26], to build trust in the B2B environment, a normative environment was used to monitor agents' activities and behaviour in MAS, dynamically collecting data for evaluation in a computational trust model to select the best partner. Negotiation facilities were also employed to find agreements between parties, culminating in electronic contracts to support the B2B trust environment by integrating the normative environment, computational trust model, and negotiation tools using the ANTE software platform.

Additionally, in [27], a method was proposed to establish trust between agents through the negotiation process using trust and sanctions. This method was tested in an experimental scenario involving betrayal during partner selection, where the truster agent evaluates the possibility of betrayal based on contract violations and sanctions and the value of relationships based on previous and current contracts. Trust is used to expect positive partner behaviour, while sanctions deter betrayal. In [26] [27], they rely on monitoring the behaviour and activity of the agent in the environment to evaluate trust after negotiation between agents.

In [28], trust is measured and evaluated through direct trust based on direct interactions between agents and recommendations from other agents, i.e., indirect trust, with trust values stored in the blockchain to ensure security.

In addition, in [29], a new method was proposed in a maze-like environment to ensure trust in MAS using interactive proof. A control strategy based on the Partially Observing Markov Decision Process (POMDP) a method that was used to monitor agents and ensure they follow the policies of their team members, and ensures that all agents work as a cohesive team and trust each other through constant verification and exchange of information. Its goal is to build trust by using interactive proofs and to confirm that all agents share the same goal and method for achieving it.

In multi-agent systems, computational trust models measure and estimate trust levels by processing data within a multi-agent environment. These models are characterized by elements such as trust dimensions (local trust, reputation, role-based trust), semantic trust, mathematical models for calculating trust or reputation, trust networks (transfer and propagation of trust among agents), as well as risk management and reliability [30]. Moreover, the decision-making process regarding agent interaction involves combining reliability, trust/reputation, and risk factors. Therefore, trust is necessary, but more is needed for decision-making under uncertainty [18]. Table 1 summarizes the main works of trust in multi-agent systems.

Table1. Works of trust in Multi-Agent Systems

Work	Technique/method Trust	Positive Points	Negative Points
[27]	Trust Model and negotiation	Establishing trust through trust and its sanctions is more positive for ensuring trust between agents than trust alone.	Lack of trust and cooperation due to sanctions leads to forced relationships.
[26]	Trust Model and negotiation	Create a trustworthy environment. Enhanced trust to reach better agreements.	Poor performance with an increasing number of agents due to its reliance on the computational trust model.
[23]	Blockchain-Smart Contract	Smart contracts are secure using cryptography techniques while being stored in a blockchain ledger. Ensures mitigation of manipulation and change by another party. Contributes to carrying out tasks.	Poor execution performance with the addition of a large number of agents in the system. Lack of flexibility because smart contracts do not adapt to new demanding changes.
[28]	Blockchain	Provides trust between agents for lack of intermediary (decentralized blockchain technology). Transparency because the calculated trust value is immutable and secure.	Difficulty correcting the error.
[25]	Decision-making	Ensure the value of trust thanks to the behaviour of agents through their behavioural history.	Not studying the security of agents' behavioural history.
[29]	Interactive Proof	Difficulty in forging sent messages. Ensure trust between agents to follow the same team policy.	The possibility of the opposing agent playing the role of the attacker, the man in the middle.
[24]	Blockchain-Smart Contract	Reducing labor and maintenance costs. Agents share more quickly. A decentralised system between vehicles and the service provider.	Changes in the working mechanism lead to a lack of flexibility, necessitating the redeployment of the smart contract.

3. Trust in Normative Multi-Agent System

Mutual trust between agents frequently contributes to forming “*deontic logic*” organisational rules, i.e. obligations [31]. For example, a group of agents in a work environment work on organising spaces. The agent ‘A’ always works on organizing space ‘y’ while the rest of the agents have other spaces. Others are also organised, and by continuing and repeating each time that agent ‘A’ is the one who organises space ‘y’, the agents have come to trust agent ‘A’ to organise space ‘y’ (collective trust), which leads to the creation of a rule or obligation that agent ‘A’ is responsible for space ‘y’ in organising and thus it must adhere to this rule.

The norms in the normative system are rules that determine the behavioural pattern of agents in the environment. They are described as basic norms in the system that affect the agents by reward and sanction [32] [33]. Consequently, it makes the action and behaviour of the agents expected in the system because it is subject to regulatory norms (i.e., deontic logic), which are permission, prohibition, and obligation [31] [32].

The obligations of agents are essentially considered goals for the normative agent by following the strategy of "your wish is my command". This later means that the desire of the norm agent must be realized by the agents and adhered to [34]. A normative agent is responsible for defining and implementing norms that agents must adhere to.

In addition, the commitments defined in the system define agent behaviour and help it identify potential agents that can contribute to achieving their goals by ensuring they behave in reliable and agreed-upon ways [33] and thus, this leads to being trusted to perform the task.

On the other hand, agents in the normative multi-agent system must adopt, comply with, and adhere to these norms. [33] Proposed the norm's trust framework to decide which norm agents should adopt. This framework contains two modules: the assessment module and the decision-making module. The assessment module consists of the **Conformer Agent Evaluation** based on the transitive trust concept. Agent A trusts Agent B (a credible agent) because it adhered to and practiced the norms. As a result, Agent A infers trust in this norm. The norms' trust assessment process was assessed regarding the adoption ratio, norm salience, and norm adoption risk.

The decision-making module uses the norms' trust (NT) value to determine the norm adoption decision depending on the readiness level: Adopt if the NT is high, Emulate if the NT is low, and ignore if the NT is null.

The decision-making module uses the norms' trust (NT) value to determine the norm adoption decision depending on the readiness level: Adopt if the NT is high, Emulate if the NT is low, and ignore if the NT is null.

Also, in [35], the authors proposed a norm's trust model, which is used to adopt or reject norms, where the norm's benefits are evaluated through authority, reputation, and the norm's adoption rate by the agents. Authority means monitoring the authorized agents in the environment; they have the authority to reward and sanction agents in society and have a high trust value. Reputation is based on neighbors' opinions and direct trust, while adoption is based on the actions of agents in the environment. After observing and detecting potential norms in the environment, the decision to adopt or reject is made according to the trust value in the criteria, either the highest possible value (1, i.e., trust) or the lowest value (0, i.e., distrust).

Moreover, [34] considers that decision-making is rational in multi-agent normative systems thanks to existing norms and penalties in the case of violation. We can distinguish strict norms, such as "you cannot enter the train station without a ticket," from easy norms, which include the possibility of riding the train without a ticket, but there is the possibility of examination and punishment. There is also a difference in the causes of violation, as there is a violation of the obligation due to its preference over punishment, the obligation conflicts with the goals and desires of the agent, and the behaviour is considered non-violative and has no punishment. Therefore, rational decision-making determines the agreement and balance between preferences and norms for agents.

Therefore, in [36], a Trust Framework for Electronic Payment Analysis based on a combination of negotiation contracts and trust models is proposed. The negotiation contract means negotiation between agents in the system to establish norms and obligations through the mental attitude, that is,

the agents' beliefs, goals, and desires, and linking them to the necessary penalties in case of violation and the addition of controls to enforce sanctions. The trust model imposes trust between the agents and cooperation, either individual, which is the trust of the parties, or control trust, which is collective and joint trust.

In conclusion, it can be observed that the primary works addressing trust in normative multi-agent systems have focused on studying how agents can choose which norms to apply while considering the trust factor [34] [35]. Consequently, the main techniques used are decision-making techniques, as an agent needs to evaluate several norms to determine the best one for its situation. And most researchers worked on methods for adopting the norm. In our research, we approach the problem from a different perspective, where we will work on the security of norms because they prevent unwanted behavior. It is widely accepted that norms can ensure trust in multi-agent systems because they regulate agents' behaviors. However, when norms are considered an essential component of a normative multi-agent system, they may become targets of various threats. In such cases, norms may lose their reputation, and trust between agents cannot be ensured through norms alone. Therefore, we believe there is a significant gap in the specialized literature regarding the question: How can norms reinforce trust between agents?

In our future research, we focus on a model where we have agents, norms, and trust as a technique to protect norms.

4. Conclusion

This paper proposes a survey about trust in normative multi-agent systems. In fact, trust is an essential concept in multi-agent systems, especially the open ones. Also, norms contribute to enhancing trust between agents. However, only some works have targeted the problem of trust in normative multi-agent systems. We present a survey about trust in multi-agent systems by describing the main works and their techniques. And studied trust in normative multi-agent systems. In fact, despite the consensus that norms represent a method to ensure the trust of agents, only a few works targeted trust in normative multi-agent systems. Moreover, almost all the works in this field propose techniques for adopting norms (i.e., How can agents trust norms?).

In the future, we propose to study this problem from another perspective. Our goal is to use norms to reinforce trust.

Bibliography

- [1] G. Boella, L. Van Der Torre, and H. Verhagen. "Introduction to normative multiagent systems". *Computational & Mathematical Organization Theory*, vol 12, pp. 71–79, 2006.
- [2] M.A. Mahmoud, M.S. Ahmad, M.Z.M. Yusoff, A. Mustapha, et al. "A review of norms and normative multiagent systems". *The Scientific World Journal*, 2014.
- [3] A. Dorri, S.S. Kanhere, and R. Jurdak. "Multi-agent systems: A survey". *Ieee Access*, vol 6, pp. 28573–28593, 2018.
- [4] E.J. De Visser, R. Pak, and T.H. Shaw. "From 'automation' to 'autonomy': the importance of trust repair in human-machine interaction". *Ergonomics*, vol 61, n°10, pp. 1409–1427, 2018.
- [5] N.H.A. Hamid, M.S. Ahmad, A. Ahmad, M.A. Mahmoud, M.Z.M. Yusoff, and A. Mustapha. "Trusting norms in normative multi-agent systems". 2015.
- [6] J. J.Ch. Meyer and R.J. Wieringa. "Deontic logic in computer science: normative system specification". *John Wiley & Sons, Inc.*, 1994.
- [7] J. Granatyr, V. Botelho, O.R. Lessing, E.E. Scalabrin, J.P. Barthès, and F. Enembreck. "Trust and reputation models for multiagent systems". *ACM Computing Surveys (CSUR)*, vol 48, n° 2, pp.1–42, 2015.

- [8] A. Aref and T. Tran. "A trust establishment model in multi- agent systems". In *Workshops at the twenty-ninth aaai conference on artificial intelligence*, 2015.
- [9] N.H.A. Hamid, M.S. Ahmad, A. Ahmad, and A. Mustapha. "A norm trust framework for norms adoption in normative multi- agent systems". *Jurnal Teknologi (Sciences and Engineering)*, vol 77, n° 33, pp. 7–16, 2015.
- [10] M.Momani and S. Challa. "Survey of trust models in different network domains". *arXiv preprint arXiv:1010.0168*, 2010.
- [11] N. Drawel, J. Bentahar, and E. Shakshuki. "Reasoning about trust and time in a system of agents". *Procedia Computer Science*, vol 109 , pp .632– 639, 2017.
- [12] D. Artz and Y. Gil. "A survey of trust in computer science and the semantic web". *Journal of Web Semantics*, vol 5, n°2, pp. 58–71, 2007.
- [13] F. Hendrikx, K. Bubendorfer, and R. Chard. "Reputation systems: A survey and taxonomy". *Journal of Parallel and Distributed Computing*, vol 75, pp. 184–197, 2015.
- [14] L. Mui, M. Mohtashemi, and A. Halberstadt. "A computational model of trust and reputation". In *Proceedings of the 35th annual Hawaii international conference on system sciences*, pp 2431–2439. IEEE, 2002.
- [15] B.G. Robbins. "What is trust? a multidisciplinary review, critique, and synthesis". *Sociology compass*, vol 10,n°10, pp. 972–986, 2016.
- [16] D. Trček and D. Trček. "A brief overview of trust and reputation over various domains". *Trust and Reputation Management Systems: An e- Business Perspective*, pp 5–19, 2018.
- [17] A. Jøsang, R. Ismail, and C. Boyd. "A survey of trust and reputation systems for online service provision". *Decision support systems*, vol 43 , n° 2, pp. 618–644, 2007.
- [18] C. Burnett, T.J. Norman, and K. Sycara. "Trust decision- making in multi-agent systems". In *Twenty-second international joint conference on artificial intelligence*, 2011.
- [19] H. Huang, G. Zhu, and S. Jin. "Revisiting trust and reputation in multi-agent systems". *ISECS International Colloquium on Computing, Communication, Control, and Management*, vo 1, pp. 424–429. IEEE, 2008.
- [20] H. Jiang, Y. Yuan, T. Guo, and P. Zhao. "Measuring trust and automatic verification in multi-agent systems". *8th International Conference on Dependable Systems and Their Applications (DSA)*, pp. 271–277. IEEE, 2021.
- [21] R. Hermoso, H. Billhardt, and S. Ossowski. "Role evolution in open multi-agent systems as an information source for trust". In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*: vol 1, pp. 217–224, 2010.
- [22] S.D. Ramchurn, D. Huynh, and N.R. Jennings. "Trust in multi-agent systems". *The knowledge engineering review*, vol 19, n°1, pp.1–25, 2004.
- [23] K. Lad, M.A.A Dewan, and F. Lin. "Trust management for multi-agent systems using smart contracts". In *IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCCom/CyberSciTech)*, pp. 414–419. IEEE, 2020.
- [24] H. Mhamdi, B.O. Soufiene, A. Zouinkhi, O. Ali, and H. Sakli. "Trust-based smart contract for automated agent to agent communication". *Computational Intelligence & Neuroscience*, 2022.
- [25] M. Cheng, C. Yin, J. Zhang, Shahin Nazarian, J. Deshmukh, and P. Bogdan. "A general trust framework for multi-agent systems". In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 332–340, 2021.
- [26] E. Oliveira, H.L. Cardoso, M.J. Urbano, and A.P. Rocha. "Normative monitoring of agents to build trust in an environment for b2b". In *Artificial Intelligence Applications and Innovations: 10th IFIP WG 12.5 International Conference, AIAI 2014, Rhodes, Greece, September 19-21, 2014*. Proceedings 10, pp.172–181. Springer, 2014.
- [27] J. Urbano, H.L. Cardoso, A.P. Rocha, and E. Oliveira. "Trust and normative control in multi-agent systems". 2012.
- [28] O. Samuel, N. Javaid, A. Khalid, M. Imrarn, and N. Nasser. "A trust management system for multi-agent system in smart grids using blockchain technology". In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6. IEEE, 2020.

- [29] J.A. Delamer and S. Givigi. "Controller-based interactive proof to ensure trust in multiagent systems". *IEEE Transactions on Automatic Control*, vol 67, n°9, pp. 4963–4970, 2022.
- [30] G. Lu, J. Lu, S. Yao, and Y.J. Yip. "A review on computational trust models for multi-agent systems". *The open information science journal*, vol 2, pp. 18–25, 2009.
- [31] C. Smith and A. Rotolo. "Collective trust and normative agents". *Logic Journal of IGPL*, vol 1, n°8 1, pp. 195–213, 2010.
- [32] M.A. Mahmoud, M.S. Ahmad, M.Z.M. Yusoff, and A. Mustapha. "A review of norms and normative multiagent systems". *The Scientific World Journal*, vol 1, pp. 684587, 2014.
- [33] N.H.A. Hamid, M.S. Ahmad, A. Ahmad, and A. Mustapha. "A norm trust framework for norms adoption in normative multi-agent systems". *Jurnal Teknologi*, vol 77, n°33, 2015.
- [34] G. Boella and L. van der Torre. "Normative multiagent systems and trust dynamics". In *International Workshop on Trust in Agent Societies*, pp 1–17. Springer, 2003.
- [35] A.M.K. Itaiwi, M.S.Ahmad, and A.Y.C. Tang. "Norm's trust model to evaluate norm benefit awareness for norm adoption in an open agent community". *International Journal of Advanced Computer Science And Applications*, vol 9, n°2, 2018.
- [36] R. Vincent, F. Adekoya, and P. Omoleh. "A normative multi- agent trust framework for electronic payment analysis". *International Journal of Computer Science Issues (IJCSI)*, vol 10, n°2, pp 346, 2013.

POSTER SESSION



Web Clicks Analysis by Episode Mining

Ouarem Oualid^{1,2✉}, Nouioua Farid^{1,2}, Mekhoukh Iamis², Hebboul Ahlem²

¹ Computer Science department, Mohamed El Bachir El Ibrahimi University of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria
mekhoukhIamis@gmail.com, hebboulahlem@gmail.com

² Laboratory of Materials and Electronic Systems (LMSE). Mohamed El Bachir El Ibrahimi University of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria oualid.ouarem@univ-bba.dz, farid.nouioua@univ-bba.dz

Abstract

Data-mining is a powerful tool for analyzing large volumes of data in order to extract useful and implicit knowledge. Recently, a new collection of data-mining techniques for temporal data has been proposed for extracting sequential patterns. This field is called frequent episode mining or simply episode mining. Recent works have proved that episode mining is a very important tool for extracting temporal patterns from a long sequence of events called episodes. However, there is a serious lack of applications that use such algorithms to analyze captured data. In this paper, we propose an application of a recent episode mining algorithm called EMDO to analyze users' web clicks and help in modeling the behavior of users using a number of rules.

Keywords: Episode Mining, sequence, distinct occurrence, episode rule

Personality trait- driven ChatGPT adoption

Rima Benfredj ¹✉, Farid Nouioua ¹, Abderraouf Bouziane ¹

¹ Department of Computer Science, University of Mohamed El Bachir El Ibrahimi Bordj Bou Arréridj, Algeria.

Laboratory L.M.S.E, University of Mohamed El Bachir El Ibrahimi, Bordj Bou Arréridj, Algeria Email of
Author 1: rima.benfradj@univ-bba.dz.

Email of Author 2: farid.nouioua@univ-bba.dz.

Email of Author 3: abderraouf.bouziane@univ-bba.dz

Abstract

In this current work, we aim to simulate the adoption of ChatGPT on social media taking into account the personality traits as a key factor; we extend our previously suggested innovation diffusion framework. Our initial model proposed four fundamental stages of innovation adoption, by using random variables for perceived innovation features and personality factors. However, in this study we improve the model using a hybrid model that combines BERT and Random Forest to predict the Big Five personality traits (Openness, Conscientious, Extraversion, Agreeableness, and Neuroticism) values from a dataset of ChatGPT-related tweets and user metadata. The simulation examines adoption rates for the two different works; the initial model with randomly assigned personality attributes and various scenarios, and the revised model with the real predicted personality values. The results reveal a considerable increase in adoption rate in the current improved model, illustrating the importance of personality-driven behaviors in innovation diffusion. This study emphasizes the impact of the hybrid model BERT and Random forest for providing realistic users' personality traits. Our findings help to better understand the psychological elements that drive the adoption of AI technologies such as ChatGPT in social networks.

Keywords: Innovation adoption, personality traits, OCEAN, ChatGPT, BERT, Random Forest.

A new heuristic for the flow-shop scheduling problem

Khaoula Djemla^{1✉}, Abdelhamid Benhocine¹, Nasser Eddine Mouhoub²

¹ Ferhat Abbas University, Setif1, Mathematical Departement, LMFN Laboratory, Setif, Algeria

{ khaoula.djemla , abenhocine }1@univ-setif.dz ²

² M'sila University, College of Mathematics and Computer Science, M'sila, Algeria

nassereddine.mouhoub2@univ-msila.dz

Abstract

Shop scheduling is a crucial area of management. It involves planning the various tasks and operations within a production environment. The main objective of shop scheduling is to a smooth and efficient production flow by finding an execution order for the different activities in order to minimize production delays and maximize workshop efficiency. There are several types of shop scheduling: single-machine scheduling, parallel scheduling, and sequential scheduling. A new heuristic is proposed to solve the sequential flow-shop scheduling problem. This heuristic is based on sorting the processing times of the different tasks on each machine, allowing for the determination of the corresponding task order.

Keywords: scheduling, heuristic, flow-shop