IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154235

Autonomous Multi-Agent Pipeline for Biomedical Research Automation

Murugeswari K[1], Pradeesh S[2], Syed Umar Nafeez G[3], Vimal M[4], Vinny Sam Francis V[5]

AIDS DSEC Perambalur Tamil Nadu India[1-5]

Abstract: This paper presents a multi-agent system for biomedical question answering that emulates a structured research workflow using large language models (LLMs). The proposed architecture leverages a LangGraph-based pipeline in which multiple specialized agents collaboratively perform literature retrieval, hypothesis generation, experimental protocol design, and validation. Three parallel junior agents gather evidence from diverse biomedical sources, followed by a supervisor that synthesizes a unified hypothesis. Subsequent agents refine the hypothesis, design experimental protocols, and conduct peer and safety reviews through an iterative feedback loop. A principal investigator agent produces a final decision, while an evaluator module acts as an LLM-as-a-judge to assess quality, precision, recall, latency, and cost. The system integrates retrieval-augmented generation (RAG), structured JSON outputs, and retry mechanisms to ensure robustness and consistency. Experimental evaluation demonstrates that the proposed approach improves reasoning depth, factual grounding, and decision reliability compared to single-agent baselines. Additionally, the framework provides transparent cost and latency tracking, making it suitable for real-world research assistance applications.

Keywords: Multi-Agent Systems, Biomedical Question Answering, Large Language Models, Retrieval-Augmented Generation, LangGraph, Experimental Protocol Design, LLM Evaluation, AI in Healthcare.

I. INTRODUCTION

A. Background and Context

Recent advancements in large language models (LLM**s)** have significantly transformed the landscape of biomedical research and question answering. Models such as GPT-4 and Gemini demonstrate strong capabilities in natural language understanding, reasoning, and knowledge synthesis. However, standalone LLMs often struggle with domain-specific reliability, interpretability, and factual grounding, particularly in biomedical contexts where accuracy is critical.

To address these limitations, retrieval-augmented generation (RAG) and multi-agent systems have emerged as promising paradigms. RAG integrates external knowledge sources such as PubMed and ClinicalTrials.gov, enabling models to produce evidence-based responses. Meanwhile, multi-agent frameworks distribute complex tasks across specialized agents, improving modularity, reasoning depth, and robustness. These approaches align with real-world scientific workflows, where multiple experts collaborate iteratively to generate and validate hypotheses.

B. Problem Statement

Despite the progress in LLMs and RAG systems, existing biomedical question answering solutions face several critical challenges. First, most systems rely on single-agent architectures, which limit their ability to perform multi-step reasoning and cross-validation. Second, retrieved information is often not systematically synthesized, leading to inconsistencies and reduced interpretability. Third, there is a lack of structured evaluation mechanisms that assess not only answer quality but also cost, latency, and reliability. Finally, current systems do not adequately replicate the iterative and collaborative nature of scientific research, resulting in shallow or incomplete outputs.

C. Objectives

The primary objective of this study is to design and implement multi-agent biomedical question answering system **that simulates a real-world research pipeline. Specifically, the system aims to:**

- Integrate multiple specialized agents for literature retrieval, hypothesis generation, protocol design, and validation
- Employ a structured pipeline using LangGraph to enable parallel processing and iterative refinement**.**
- Develop an evaluation module that measures performance in terms of quality, precision, recall, cost, and latency.
- Provide transparent, structured outputs that support decision-making in biomedical research.

D. Significance of the Study

This study contributes to the advancement of AI-driven research systems by introducing a scalable and modular multi-agent architecture tailored for biomedical applications. The proposed approach enhances **reasoning reliability**, **factual**

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154235

grounding, and **decision transparency**, addressing key limitations of existing LLM-based systems. Furthermore, by integrating evaluation metrics such as cost and latency, the system provides practical insights for real-world deployment.

The significance of this work extends to multiple domains, including healthcare, pharmaceutical research, and clinical decision support. By automating complex research workflows and enabling evidence-driven insights, the proposed system has the potential to accelerate scientific discovery, reduce manual effort, and improve the overall efficiency of biomedical research processes.

II. LITERATURE REVIEW

A. LLM-Based Biomedical Question Answering

Early biomedical QA systems relied on rule-based or statistical methods, which were limited in handling complex queries. With the introduction of transformer-based architectures such as BERT and its biomedical variants like BioBERT, significant improvements were achieved in domain-specific understanding.

More recently, advanced LLMs such as GPT-4 and Gemini have demonstrated strong reasoning and generative capabilities. These models can interpret complex biomedical queries and generate coherent responses. However, they are prone to hallucinations and lack consistent grounding in verified scientific evidence, which limits their reliability in critical applications.

B. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has emerged as an effective approach to mitigate hallucination by incorporating external knowledge sources into the generation process. Systems leveraging databases such as PubMed and ClinicalTrials.gov enable models to access up-to-date and domain-specific information.

RAG frameworks typically involve document retrieval, embedding-based similarity search, and context-aware response generation. While this improves factual accuracy, existing implementations often lack structured reasoning and fail to integrate insights from multiple sources effectively. Additionally, most RAG systems operate as single-agent pipelines, limiting their ability to perform iterative refinement or cross-validation of results.

C. Multi-Agent Systems in AI Research

Multi-agent systems have gained attention as a means to enhance problem-solving by distributing tasks among specialized agents. Frameworks such as LangChain and LangGraph enable the design of collaborative AI pipelines where agents perform distinct roles such as retrieval, reasoning, and validation.

Recent studies demonstrate that multi-agent architectures can improve reasoning depth, modularity, and robustness compared to monolithic models. By simulating human workflows—such as collaboration between researchers, reviewers, and decision-makers—these systems enable iterative refinement and error correction. However, many existing implementations lack domain-specific customization for biomedical applications and do not incorporate comprehensive evaluation mechanisms.

D. Evaluation of LLM Systems

Evaluating LLM-based systems remains a challenging task due to their generative nature. Traditional metrics such as accuracy and F1-score are insufficient for capturing reasoning quality and factual correctness. The concept of **LLM-as-a-judge** has been introduced, where models evaluate outputs based on criteria such as coherence, relevance, precision, and recall.

Despite its advantages, this approach introduces concerns regarding bias and consistency. Furthermore, most existing evaluation frameworks do not consider system-level metrics such as latency, token usage, and cost, which are critical for real-world deployment.

III. DATASET AND PREPROCESSING TECHNIQUES

A. Data Sources

The proposed system relies on heterogeneous biomedical data sources to ensure comprehensive and evidence-based reasoning. Primary data is retrieved dynamically from authoritative repositories such as PubMed, ClinicalTrials.gov, and preprint platforms including bioRxiv and medRxiv. Additional regulatory and global health information is incorporated from sources such as World Health Organization and U.S. Food and Drug Administration.

Unlike traditional machine learning systems that rely on static datasets, this approach adopts a **dynamic retrieval paradigm**, where relevant documents are fetched in real time based on the input query. This ensures that the system operates on up-to-date and context-specific biomedical knowledge, which is essential for research-oriented applications.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154235

B. Data Collection Strategy

Data collection is performed through a retrieval-augmented mechanism using a search-based API. Each query is processed by specialized agents that generate search queries and iteratively retrieve relevant documents. The system employs a **ReAct (Reasoning + Acting)** framework, enabling agents to refine search queries across multiple iterations and gather diverse evidence.

To maintain data quality, domain-specific constraints are applied to restrict retrieval to trusted biomedical sources. Duplicate URLs are filtered, and summarized search results are prioritized to reduce redundancy. This structured retrieval strategy ensures that only high-quality and relevant information is passed to downstream components.

C. Preprocessing Techniques

Raw text is cleaned to remove HTML tags, special characters, and irrelevant metadata. Standard normalization techniques such as lowercasing and whitespace standardization are applied to ensure consistency. Long documents are divided into smaller, semantically coherent chunks using recursive text splitting methods. This improves retrieval efficiency and enables the model to process context within token limits. Text chunks are transformed into dense vector representations using embedding models. These embeddings capture semantic similarity, enabling efficient retrieval of relevant information during query processing. Retrieved chunks are ranked based on relevance scores, and only the top-k most relevant segments are selected. This step reduces noise and ensures that the model focuses on high-quality evidence. The processed data is converted into structured formats (e.g., JSON schemas) to ensure compatibility with downstream agents. This enables consistent interpretation and reduces ambiguity in multi-agent communication.

IV. METHODOLOGY

A. *SYSTEM ARCHITECTURE DESIGN*

The system follows a directed acyclic graph (DAG) with iterative feedback loops, where each node represents a distinct agent responsible for a specific task. The pipeline begins with parallel evidence collection and converges into a centralized reasoning and validation process.

Three junior agents operate concurrently to retrieve biomedical information from diverse sources. Their outputs are aggregated by a supervisor agent, which synthesizes a unified hypothesis. This hypothesis is iteratively refined and validated through multiple stages, including protocol design, peer review, and safety analysis. The workflow concludes with a principal investigator agent that produces the final decision, followed by an evaluation module that assesses system performance.

B. Multi-Agent Workflow Design

Three specialized agents independently retrieve information from domain-specific sources such as PubMed and ClinicalTrials.gov. Each agent uses a ReAct-based search strategy to iteratively refine queries and gather relevant evidence. This parallelization improves coverage and reduces retrieval bias.The supervisor agent aggregates outputs from all junior agents and generates a coherent, unified hypothesis. This step resolves conflicts, removes redundancies, and ensures logical consistency across multiple evidence sources.A dedicated agent enhances the synthesized hypothesis by incorporating detailed attributes such as molecular targets, biomarkers, and dosage considerations. This stage increases the specificity and scientific relevance of the hypothesis.The protocol designer agent generates a structured experimental plan, including study type, cohort selection, intervention strategy, endpoints, and estimated cost. This transforms abstract hypotheses into actionable research procedures.Two independent agents—peer reviewer and safety officer—evaluate the proposed protocol. The peer reviewer focuses on scientific rigor and validity, while the safety officer assesses ethical and clinical risks. Their outputs include severity levels (e.g., CRITICAL, MAJOR, MINOR). Based on the review severity, the system dynamically routes the workflow back to either the hypothesis refinement or protocol design stage. This loop continues for a predefined number of iterations, ensuring progressive improvement and error correction. The final output is evaluated using an LLM-based scoring mechanism that assesses quality, coherence, precision, recall, and factual grounding. Additionally, system-level metrics such as latency, token usage, and cost are recorded for performance analysis.

C. Retrieval-Augmented Generation Integration

The system incorporates retrieval-augmented generation (RAG) to ensure evidence-based reasoning. Retrieved documents are processed, ranked, and injected into the context of each agent. This approach reduces hallucinations and improves factual accuracy by grounding responses in verified biomedical literature.

D. Data Flow and State Management

The pipeline maintains a shared state that stores intermediate outputs, token usage, and latency metrics for each agent. This state is incrementally updated as the workflow progresses, enabling traceability and reproducibility. Structured data exchange using JSON schemas ensures consistency and minimizes ambiguity across agents.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154235

E. Robustness and Error Handling

To enhance reliability, the system employs multiple safeguards, including retry mechanisms for failed responses, validation of structured outputs, and controlled iteration limits. These measures ensure stable execution and consistent performance, even in complex query scenarios.

V. SYSTEM ARCHITECTURE

A. Architectural Overview

The system follows a directed graph architecture consisting of interconnected nodes, where each node represents a specialized agent. The workflow begins with parallel data retrieval and progresses through synthesis, refinement, validation, and final decision-making stages. Accepts a biomedical research question from the user. Executes the multi-agent pipeline, including retrieval, reasoning, and validation.Manages workflow routing, iteration loops, and state transitions.Produces the final report along with evaluation metrics such as cost, latency, and quality scores.

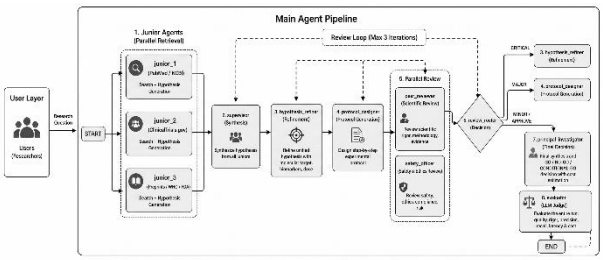

Fig 1: system architecture

B. Core Components

The system accepts natural language queries related to biomedical research. These queries serve as the initial input for the pipeline and trigger the execution of all downstream processes. Three independent agents perform concurrent information retrieval from trusted biomedical sources such as PubMed and ClinicalTrials.gov. This module ensures diverse and comprehensive evidence collection. The supervisor aggregates outputs from all retrieval agents and generates a unified hypothesis. It acts as a central reasoning unit that resolves inconsistencies and ensures coherence. This component enhances the synthesized hypothesis by incorporating domain-specific attributes such as molecular targets, biomarkers, and dosage parameters, improving scientific precision. The protocol designer converts the refined hypothesis into a structured experimental plan, including study design, intervention methods, evaluation metrics, and estimated costs. A dedicated routing component dynamically directs the workflow based on review outcomes. If critical issues are identified, the system loops back to earlier stages (hypothesis refinement or protocol design). This iterative mechanism ensures continuous improvement. The principal investigator synthesizes all validated outputs and produces a final decision, including justification, risk assessment, and recommendations. The evaluator implements an LLM-as-a-judge approach to assess output quality, precision, recall, and coherence. It also computes system-level metrics such as token usage, latency, and cost.

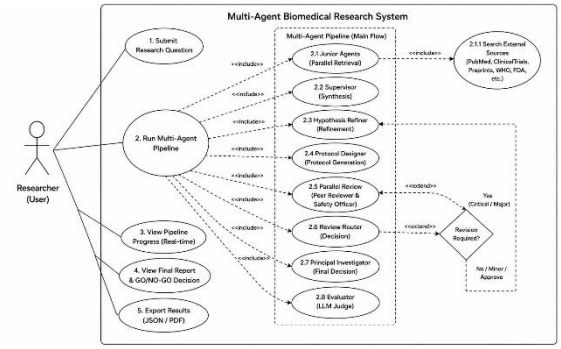

Fig2: Use Case Diagram

C. *Data Flow and Communication*

All components communicate through a shared state object, which stores intermediate results, metadata, and performance metrics. Structured data exchange using JSON schemas ensures consistency and reduces ambiguity between agents. The

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154235

architecture supports both fan-out (parallel execution) and fan-in (aggregation) patterns, enabling efficient data processing.

D. ***Parallelism and Scalability***

The use of parallel retrieval and review modules improves system efficiency and scalability. By distributing tasks across multiple agents, the architecture reduces latency and enhances coverage. Additional agents can be integrated without affecting the overall structure, making the system extensible for future enhancements.

E. ***Reliability and Monitoring***

The system incorporates retry mechanisms, validation checks, and controlled iteration limits to ensure robustness. Performance metrics such as latency and cost are tracked at each stage, providing transparency and enabling optimization.

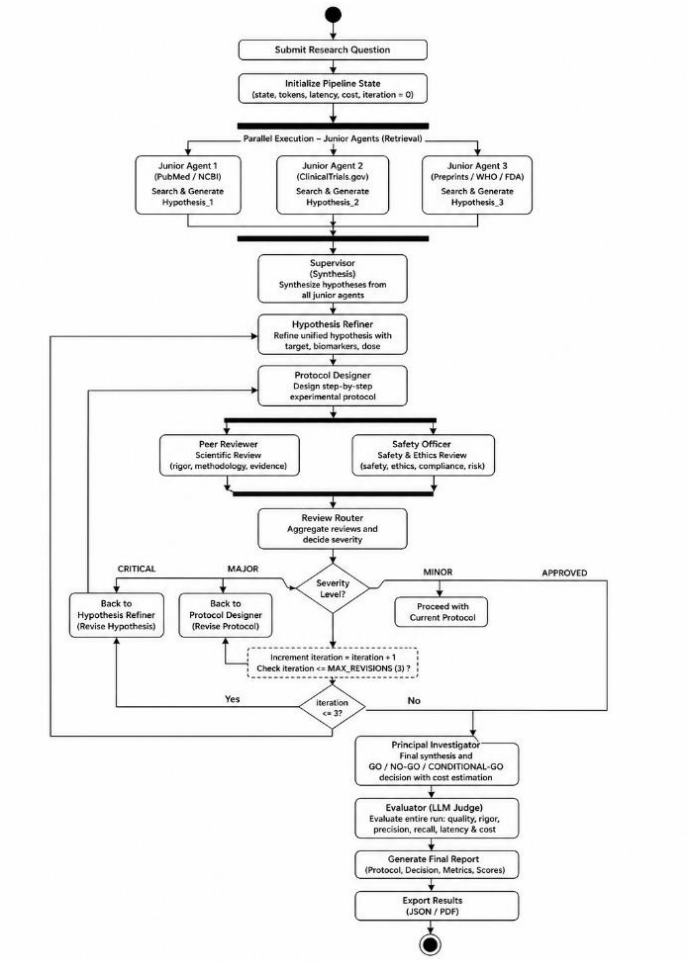

Fig 3: Activity Diagram

VI. RESULTS

A. Experimental Setup

The system was evaluated using a set of biomedical research questions covering domains such as drug repurposing, disease treatment, and clinical trial analysis. The pipeline was executed end-to-end using the implemented multi-agent architecture in LangGraph, with retrieval performed from trusted sources including PubMed and ClinicalTrials.gov. Performance was assessed using both qualitative and quantitative metrics**,** including Answer relevance and coherence, Precision and recall, Factual grounding, and Latency

B. Qualitative Results

The system demonstrated strong performance in generating structured, research-oriented outputs. Unlike traditional single-agent systems, the proposed approach produced Clearly defined hypotheses with supporting evidence, Detailed experimental protocols including dosage, cohorts, and endpoints, Explicit risk assessments and safety considerations, Final decisions (GO/NO-GO) with justifications.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154235

The iterative review mechanism significantly improved output quality. In multiple test cases, initial hypotheses were refined through feedback loops, resulting in more precise and scientifically valid conclusions. The integration of retrieval ensured that responses were grounded in credible biomedical literature, reducing hallucinations..

C. Quantitative Results

The LLM-based evaluator indicated high scores in coherence, relevance, and completeness. Precision improved due to domain-restricted retrieval, while recall benefited from parallel evidence collection by multiple agents. The total pipeline latency was influenced by the number of agents and iteration cycles. Parallel execution of retrieval and review stages reduced overall processing time compared to sequential approaches. However, additional review loops increased latency in complex cases.Token usage was tracked at each stage, enabling precise cost estimation. The results show that reasoning-intensive agents contributed the most to cost, while retrieval agents remained relatively lightweight. Despite higher computational overhead than single-agent systems, the improved output quality justifies the cost in research-critical applications. The system maintained stable performance across different query types. The modular architecture allowed efficient scaling by adding or modifying agents without disrupting the pipeline.

VII. COMPARATIVE ANALYSIS

A .Comparison with Single-Agent LLM Systems

Traditional LLM-based systems, such as GPT-4 and Gemini, rely on a single model to perform retrieval, reasoning, and response generation. While these systems demonstrate strong generalization and fluency, they exhibit several limitations in biomedical contexts Complex multi-step problems are often simplified or partially addressed, Lack of grounding leads to generation of unsupported or incorrect information, Outputs are typically unstructured and lack transparency. In contrast, the proposed system distributes tasks across specialized agents, enabling deeper reasoning, structured outputs, and improved reliability**.**

B. Comparison with RAG-Based Systems

RAG-based systems enhance LLMs by integrating external knowledge sources such as PubMed and ClinicalTrials.gov. These systems improve factual accuracy by grounding responses in retrieved documents. However, most RAG implementations follow a linear pipeline, where retrieval is followed by a single generation step. This approach has the following limitations No mechanism to revise outputs based on feedback, Difficulty in combining insights from multiple sources effectively, Absence of peer or safety review processes. The proposed system extends RAG by embedding it within a multi-agent, iterative workflow**,** enabling continuous refinement and cross-validation of results.

C. Comparison with Existing Multi-Agent Systems

Recent frameworks such as LangChain and LangGraph support the development of multi-agent pipelines. While these frameworks provide modularity and flexibility, many existing implementations Are general-purpose and not tailored for biomedical applications, Lack domain-specific retrieval constraints**,** Do not incorporate comprehensive evaluation metrics. The proposed system differentiates itself by integrating domain-restricted retrieval, structured agent roles, and an evaluation module that measures both qualitative and quantitative performance.

VIII . CONCLUSION

A. Summary of the Study

This study presented a multi-agent biomedical question answering system that integrates retrieval-augmented generation with structured agent collaboration to emulate a real-world research workflow. The system leverages LangGraph to coordinate multiple specialized agents responsible for literature retrieval, hypothesis generation, protocol design, validation, and final decision-making. By incorporating parallel processing, iterative refinement, and an LLM-based evaluation mechanism, the proposed approach addresses key limitations of traditional single-agent and linear RAG systems, resulting in improved reasoning depth, factual accuracy, and interpretability.

B . Key Outcomes and Contributions

The experimental results demonstrate that the proposed system significantly enhances the quality and reliability of biomedical question answering. The integration of domain-specific retrieval from sources such as PubMed and ClinicalTrials.gov ensures strong factual grounding, while the multi-agent workflow enables effective synthesis and validation of information. The iterative review mechanism further improves output precision and coherence. Although the system introduces additional computational cost and latency, these trade-offs are justified by the substantial gains in performance and decision reliability.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154235

C. Future Scope and Improvements

Future research can focus on enhancing the system by incorporating more advanced domain-specific models, expanding the range of biomedical data sources, and improving retrieval efficiency. Further optimization of latency and cost through model compression and adaptive agent selection can make the system more suitable for large-scale deployment. Additionally, integrating human-in-the-loop validation and extending the framework to other domains such as legal or financial research can broaden its applicability and impact.

REFERENCES

[1]. S. Zhang, Y. Chen, and X. Liu, "A Multi-Agent System for Medical Decision Support Using Artificial Intelligence," 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkey, 2023, pp. 2100-2105, doi: 10.1109/BIBM58861.2023.10385621.

[2]. H. Wang, S. Li, S. Cao, R. Yang, J. Zeng, Z. Qian, and X. Zhang, "On physicallyoccluded fake identity document detection," in Proc. 31st ACM Int. Conf. Multimedia. New York, NY, USA: Association for Computing Machinery, Oct. 2023, p. 1556.

[3]. L. Zuo, W. Chen, Q. Hong, L. Huang, Z. Wang, and Y. Chen, "An intelligent knowledge extraction framework for recognizing identification information from real-world id card images," IEEE Access 7, 2019.

[4]. Tropin, Daniil V. et al. "Improved algorithm of ID card detection by a priori knowledge of the document aspect ratio." International Conference on Machine Vision (2021).

[5]. Pratama, M. Octaviano et al. "Indonesian ID Card Recognition using Convolutional Neural Networks." 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (2018): 178-181.

[6]. B Benalcazar, Daniel P. et al. "Synthetic ID Card Image Generation for Improving Presentation Attack Detection." IEEE Transactions on Information Forensics and Security 18 (2022): 1814-1824

[7]. Satyawan, Wira et al. "Citizen Id Card Detection using Image Processing and Optical Character Recognition." Journal of Physics: Conference Series 1235 (2019):

[8]. Liem, Hoang Danh et al. "FVI: An End-to-end Vietnamese Identification Card Detection and Recognition in Images." 2018 5th NAFOSTED Conference on Information and Computer Science (NICS) (2018): 2022.

[9]. Jian Zhu, Hanjie Ma, Jie Feng, Leiyan Dai; ID card number detection algorithm based on convolutional neural network. AIP Conference Proceedings 18 April 2018

[10]. Sebastian Gonzalez; Andres Valenzuela; Juan Tapia, Hybrid Two-Stage Architecture for Tampering Detection of Chipless ID Cards, IEEE Transactions on Biometrics, Behavior, and Identity Science, 2637-6407, 2021.

[11]. Reuben P. Markham; Juan M. Espín López; Mario NietoHidalgo; Juan E. Tapia, Open-Set: ID Card Presentation Attack Detection Using Neural Style Transfer, IEEE Access, 2169-3536, 2024.

[12]. Ashwini Zinjurde and Vilas Kamble, "Credit Card Fraud Detection and Prevention by Face Recognition", International Conference on Smart Innovations in Design Environment Management Planning and Computing (ICSIDEMPC), 2020.

[13]. Alharbi, F., Alshahrani, R., Zakariah, M., Aldweesh, A. and Alghamdi, A.A., 2023. YOLO and Blockchain Technology Applied to Intelligent Transportation License Plate Character Recognition for Security. Computers, Materials & Continua, 77(3).

[14]. Patil, S., Meshram, D., Bohra, M., Daulat, M., Manwatkar, A. and Gore, A., 2023, April. Enhancing Surveillance and Face Recognition with YOLO-Based Object Detection. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 373-383). Singapore: Springer Nature Singapore.

[15]. Haque, M., Faisal, S.M., Islam, M.T. and Akash, T.H., 2023. Computer Vision-Based Intelligent Classroom Systems for Efficient Power Management in Large Educational Institutions

