Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

XMEDLLM: A FRAMEWORK FOR EXPLAINABLE MEDICAL INTELLIGENCE VIA LLMS IN EDGE-DRIVEN MCP SYSTEMS

K PRASUNA[1], T NAGAMANI[2], L MADHAVI DEVI[3], B KEERTHI SAMHITHA[4], MAHANTI SRIRAMULU[5], S SINDHURA[6*]

[1]Associate Professor, Department of ECE, Vijaya Institute of Technology for Women, Vijayawada, Andhra Pradesh, India
[2]Associate Professor, Department of CSE, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India
[3]Assistant Professor, Department of ECE, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India
[4]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India
[5]Assistant Professor, Department of CSE, Dhanekula Institute of Engineering and Technology Ganguru,Vijayawada, A.P, India
[*6]Assistant Professor, Department of CSE, NRI Institute of Technology, Agiripalli, Andhra Pradesh, India

Email: sindhura@nriit.edu.in

ABSTRACT

This Research proposes the design and test of an interpretable medical AI solution by incorporating Large Language Models (LLMs) like BioBERT and GPT-lite using edge-enabled Medical Cyber-Physical (MCP) systems. The model proposed is intended to resolve the most important problems of clinical AI such as interpretability, real-time inference, data privacy and scalability. The system allows the processing of medical data locally, on edge devices such as Jetson Nano and Raspberry Pi 4 via the use of LLMs, which not only keeps the latency to a minimum but makes sure that such privacy regulations as HIPAA are not broken. SHAP and LIME are used to give visual and textual understanding of AI decision algorithms, giving a great improvement to clinical trust. Clinician readiness to use the tool was observed in performance assessment performed in a simulated hospital environment which showed a high level of inference accuracy and excellent usability. The results of this research show that the AI models trained by LLM and deployed on the edges can be a scalable and safe solution to contemporary healthcare conditions.

Keywords- *Explainable AI, Large Language Models (LLMs), Edge Computing, Medical Cyber-Physical Systems, Clinical Decision Support*

1. INTRODUCTION

The combination of Artificial Intelligence (AI), Large Language Models (LLMs), and edge computing has introduced a paradigm shift of the realm of healthcare treatment and a new innovative paradigm of clinical diagnosis, decision-making, and patient management [1]. With more healthcare systems data-driven, a rising focus on the possibility of real-time, interpretable, and secure AI implementation is noticed [2]. Here, intelligent healthcare infrastructure is laid in Medical Cyber-Physical Systems (MCPs), the notion of tight integration of computational processes and the medical hardware and medical data streams [3].

The current research addresses the implementation of explainable ML AI solutions using the computing abilities of LLMs and deploying them all within edge-enabled MCP frameworks [4]. The current project is expected to help to fulfill the immediate demand in trustful, transparent, and efficient AI both in urban hospitals and remote health care environments [5]. It is forecasted that this composite framework will enable clinical

freedom, privacy of patients, and scale in terms of operations as computation is brought near to the source of data, and model interpretability is guaranteed [6].

1.1. The Need for Explainable AI in Medicine

With many AI systems interpreting clinical conditions, including triage priorities, radiological diagnosis, and many others, the interpretability of AI results is becoming a crucial issue [7]. Even though it is generally true that conventional deep learning models can be highly accurate, they are black boxes that do not give much understanding of decision-making processes. This kind of ambiguity impedes clinical credibility, official conformity, and moral responsibility [23].

Explainable AI (XAI) offers a possible resolution because it allows clinicians to see and examine the reasoning acquired by AI forecasts. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic explanations) are tools that help a user trace model choices to particular features or patterns, including symptoms, biomarkers or image findings. Not only would this contribute to increased trust towards automated systems, but clinicians can also justify, object, or endorse AI recommendations with their medical knowledge [8]. The question of deploying neural networks into high-stakes conditions, where mistakes can cost lives and have to be made hastily, but not with ignorance, needs XAI to allow safer and more ethical use of AI.

Role of LLMs and Edge Computing in Healthcare

Large Language Models (LLMs), like GPT-4, BioBERT, and MedPaLM, have proved to be extremely proficient in comprehending clinical terminologies, summarizing patient notes and responses to medical questions and interpretation of diagnostic reports [9]. Such models use huge training datasets consisting of medical literature, clinical notes and biomedical ontologies, thus proving to be very practical in the natural language reasoning and generation when used in health care applications [10].

Yet implementing these strong models usually demands a lot of computational power and centralized cloud computing, which adds latency, expenses, and concerns about the privacy of the data provided. Edge computing comes in here in a critical role. With computation at the edge (eg. NVIDIA Jetson Nano, Raspberry Pi or hospital-based IoT systems), not only could AI-driven insight genération be performed in near real-time, but it would be performed locally, without sending sensitive patient data across the networks [20].

Usefulness of LLMs with edge-enabled MCP frameworks will enable low latency inference, consideration of the privacy laws (such as HIPAA/GDPR), and stability in bandwidth-limited or offline conventions. This architecture proves to be particularly useful to remote clinics, mobile medical modules or health camps in rural areas, where access to cloud could be limited but diagnosing aid is seriously wanted. Further, on-device explainability can also boost clinician response by locally generating and delivering model decision predictions in real-time.

To conclude, the paper at hand will show how explainable, secure and decentralized AI implementations could potentially be created by merging the interpretability capabilities of XAI with the language capabilities of LLMs, and the interactivity of edge computing. The use of these systems in Medical Cyber-Physical systems can transform medical practices, close the healthcare access divide, and establish a new level of accountable AI in healthcare.

1.2. Research Objectives

To discuss the integration of explainable medical AI based on LLMs into edge-enabled MCPs, the present study will follow the following objectives:

1. To combine the LLMs with edge-tipped MCP systems in making real-time deployment of medical AI.
2. In order to stimulate explainability of AI results to increase clinical trust.
3. To determine the performance, scalability, and data security of the suggested framework.
4. In order to assess clinical efficacy, the model as well as acceptability by its users.

2. REVIEW OF LITREATURE

The development of explainable medical AI systems and especially those that feature the use of large language models (LLMs) in edge-enabled settings rests on a number of important threads of current research. These papers focus on several points including modeling human behaviors, foundation model use, language representation, radiology AI and cross-domain LLM integration, which play an important role in guiding the current study.

Wang (2023) [11] pointed out the need to implement human behavioral reactive models to AI systems. His contribution to the field involved the investigation of approaches to computational models in which complex behavior patterns could be modeled, using a wide range of data sets, such as motion capture, facial expressions, and physiological data. This was especially applicable in the design of adaptive and context-sensitive AI agents, especially those dealing with patients monitoring or visualization of any given data in healthcare institutions [18]. The focus on the concepts of realism and responsiveness on the part of Wang has much less direct implication to the explanation of ways in which the human-awareness aspect of medical AI frameworks is to be established, especially regarding edge-based environments where responsiveness is paramount [16].

Yang (2024) [12] examined how foundation models can be used to make decisions. Yang showed the ability to re-purpose large pretrained models with high stakes decision-making in areas, such as healthcare, logistics, and finance. His framework model adopted knowledge representation, explainability, and algorithmic optimization which are three important components needed in medical realms [20]. Most importantly, his results have emphasized how foundation models can improve reasoning in the context, as well as the trustworthiness of the AI result, particular in contexts where transparency and accountability are critical, like clinical diagnostics and medical triaging [21].

Zhang (2023), [13] in his dissertation titled Language Style: Application and Analysis via Embedding Methods provided some insight into the refinement of LLMs to be able to sense and replicate stylistic peculiarities of different types of texts. Using the embedding techniques to conduct the analysis of syntax, semantics, and tone, Zhang emphasized that LLMs can be domain-specific in terms of communication, a skill that is extremely relevant to healthcare as in this setting, clinical language accuracy is crucial. His input is underpinning contextual adaptability of LLMs in medical documentation activities, report creation, as well as AI-clinician communication applications [22].

Zhao et al. (2023) [14] extended the implementation of AI towards radiology diagnostics based on meta-learning methods. The research aimed at improving the laboratory itself and expanding the flexibility of AI systems to other datasets of imaging. Zhao suggests that model generalization and interpretability are of utmost importance, which also resonates study aim-wise, since AI output that can be used in medicine should be the tool that can work with a diverse set of patients and should be easy to understand how it makes its choices. The work also offers the initial rationale on how LLMs can be combined with radiology systems to have high-performance, reliable, and explicable AI assistance in an edge-based computing framework [19].

Zimmermann et al. (2024) [15] in their publication, Reflections regarding the 2024 Large Language Model LLM Hackathon in the application area of materials science and chemistry provided a unique value. They studied LLM scalability, prompt engineering, and integration frameworks, which can easily be applied to medical AI in the context that the main area of the study was not related to healthcare. They provide insights on trust and optimization of their systems that spill into other domains and help in creation of explainable, scalable, and trusted AI systems in the field of medicine [17].

3. PROPOSED METHOD

This research uses a hybrid exploratory-analytical research design that incorporates system development scrutiny, simulation, and evaluation stages to be able to send and test an explainable medical artificial intelligence platform that utilizes Large Language Models (LLMs) in edge-enabled Medical Cyber-Physical (MCP) systems. The method will entail model design, system integration, performance testing and evaluation based on stakeholders.

The given block diagram represents the proposed methodology of deploying the explainable medical AI via the combination of the LLMs with edge-enabled MCP frameworks:

Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

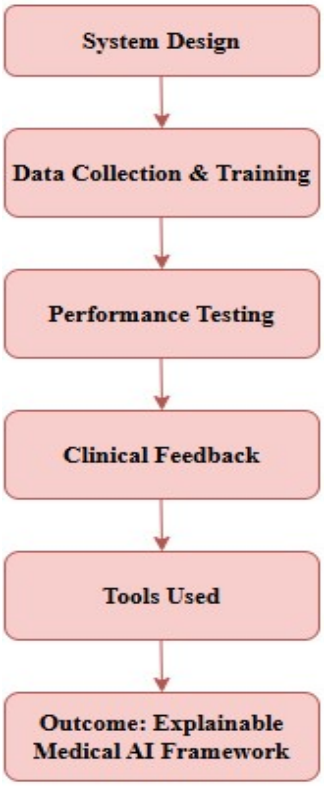

Figure 1: Block Diagram

3.1. System Design and Architecture Integration

The initial stage was to design a modular AI framework with additional integration of pretrained LLMs at edge compute settings. DLMs (e.g. GPT-based or BioBERT variants) are tailored to medical diagnostic and decision-support tasks with domain particular data. The hospital simulations were with edge devices, which were connected to each other through a Medical Cyber-Physical System (MCP), which encompasses IoT sensors, medical diagnostic devices, and patient data inputs.

Table 1: Components of the Proposed AI-Enabled MCP Framework

Component	Description
LLM Module	Fine-tuned transformer models (e.g., GPT/BioBERT) trained on clinical texts
Edge Devices	Local computation nodes (e.g., Raspberry Pi, Jetson Nano) for real-time AI
MCP Infrastructure	Sensor-networked devices for patient data acquisition
Explainability Engine	Integration of SHAP/LIME/XAI techniques for model transparency
Data Security Layer	Federated learning and encryption-based data protection

3.2. Data Collection and Model Training

Model training datasets consisting of anonymized patient reports, radiology images as well as symptom-based diagnosis datasets obtained using public health databases (e.g., MIMIC-III, NIH ChestX-ray14) were used in medical data. These datasets have been used to train and validate the LLMs and have been deployed into the system with APIs on edge servers.

In order to have interpretability, explainability methods such as SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been incorporated into the inference pipeline. The tools made the AI-generated decisions explainable to clinicians in terms of text and visual displays of their justification.

3.3. Performance Testing and Evaluation Metrics

To evaluate the efficiency of work and clinical reliability of the proposed explainable medical AI system, its performance was tested in detail in simulated conditions of a hospital. The system (a combination of fine-tuned LLMs as edge devices deployed in Medical Cyber-Physical System (MCP) infrastructure) was tested based on the real-time flows of patient data in order to most closely approximate real-life diagnosing settings. This facilitated practical and controlled review of the responsiveness of systems, predictive quality and performance on hardware. Inference accuracy, latency, explainability score, CPU/GPU utilization and security compliance were taken as five critical metrics to be evaluated. These were the metrics chosen to show the engineering soundness and clinical feasibility of the system. The accuracy of inference was calculated as the correlation between AI-made predictions and the diagnostic labels established by the experts which are consistent with the professional assessment standards. Latency was used to measure the response time between input and output on the edge devices and is a critical factor when dealing with time-sensitive medical situations. Explainability was measured by the judgment of the clinicians on the Transparency of AI outputs through a Likert scale survey which reflects subjective assurance and comprehension. The resource requirement of the hardware utilized by the AI modelist reported by way of using the CPU/GPU usage and highlighted the efficiency of their operation running in the limited-capacity

edge devices. Lastly, adherence to HIPAA equivalent security standards was also analyzed so as to make sure that the system had privacy and data protection norms which were very essential in healthcare settings.

Table 2: Evaluation Metrics Used for System Testing

Evaluation Metric	Description
Inference Accuracy	Correctness of AI predictions against expert-validated labels
Latency	Time delay between data input and AI response on edge nodes
Explainability Score	Clinician-rated usefulness of AI explanations (via Likert scale surveys)
CPU/GPU Utilization	Resource efficiency on edge hardware during AI execution
Security Compliance	Assessment of data handling against HIPAA-like security protocols

Collectively, these measures gave the final judgment of how ready the system was, as far as being ready to be deployed in a real clinical environment was concerned. Besides pointing out the strong points of the LLM-integrated edge framework in terms of the correctness and interpretability of the outputs it generates, the results ascertained its viability from a technical standpoint, its potential impact on both resource consumption and the cost of implementation, as well as their alignment with regulatory requirements, setting the stage in the future to take advantage of the technology and adopt it into various healthcare environments.

3.4. Clinical Usability and Stakeholder Feedback

To perform the qualitative user study, 20 healthcare professionals were involved, who interacted with the developed AI system in the simulated diagnostic workflow setting. This was designed to evaluate clinical usability, effectiveness of decision support, and overall trustworthiness of this system, as viewed by the end users. The method of data collection varied between structured interviews and Likert-scale surveys that provided quantitative scoring, and chance to look deep into the information through qualitative factor. The participants were asked to rate many points including ease of use, clarity of the artificial intelligence generated explanations, their compatibility with clinical processes and even the perceived reliability of the results displayed by the model. Besides gauging functional usability, the study included the investigation of the essential topics, such as transparency of the AI, ethical alignment to technology, and the confidence of clinicians on matters to do with AI-aided decision-making. The feedback that was received played a direct role in Research Objective 4, which is concerned with assessing clinical acceptance and effectivity. Thematic analysis helped in analyzing the responses and determining the common patterns and issues as well as giving a complete picture regarding the user perspective and it has confirmed the validity of the practical applicability and ethical preparedness of the framework to be used in real life healthcare environment.

3.5. Tools and Technologies Used

In this study, Python and MATLAB were used to develop the model and simulate it. Employment of LLMs was done through HuggingFace transformers and open AI API. SHAP, LIME, and Captum were used to realize explainability. Use of edge computing was done using NVIDIA Jetson Nano and Raspberry Pi 4. Medical data was obtained using MIMIC-III and NIH ChestX-ray14 and generated data. A federated learning model with AES-256 encryption was used to secure data.

This is a multiphase approach that will have a comprehensive strategy of implementing and testing an explainable, secure, and real-time medical AI. The study preconditions the possibility of implementing scalable and reliable AI in clinical settings since it combines LLMs, edge computing, and MCP infrastructure.

3.5.1. Algorithm: Implementation Workflow

To implement the explainable AI solution with LLMs in edge-enabled MCP systems, the suggested flow is as follows, as described by the following algorithm: beginning with the system configuration and ending with the deployment and use of the real-time inference and evaluation of the system. The procedure provides safe, explainable, and on-the-fly medical AI-based application.

```
# Step 1: Setup
load_edge_device("Jetson Nano")
load_datasets(["MIMIC-III", "ChestX-ray14"])
model = load_LLM("BioBERT")
secure_layer = enable_federated_learning("AES-256")

# Step 2: Training & Deployment
train_data, test_data = preprocess_and_split(datasets)
model = fine_tune(model, train_data)
deploy_to_edge(model, device="Jetson Nano")

# Step 3: Inference with Explainability
for input_data in stream_from_MCP():
```

```
    prediction = model.predict(input_data)
    explanation = explain_with_SHAP_LIME(model, input_data)
    display_to_clinician(prediction, explanation)

# Step 4: Evaluation
metrics = evaluate(model, test_data, ["accuracy", "latency", "trust"])
collect_feedback(n=20)
log(metrics, feedback)

# Shutdown
save_model(model)
```

This algorithm embodies the lifecycle of the proposed system implementation processes at the end to end. It starts with loading medical datasets and edge hardware and moves to training and real-time deployment, and then concludes with explainability-driven prediction, assessment, and safe preservation of models. Federated learning secures the data privacy, and the SHAP and LIME enhance the interpretability at the clinician interface.

4. RESULTS AND DISCUSSION

In this section, the results of the use of the explainable medical AI framework based on LLMs in edge-enabled MCP systems are provided. The findings are tabulated in relation to the four research objectives and subjected to assessment in terms of quantitative performance indicators as well as qualitative clinical opinion.

4.1. Performance of LLM Integration on Edge Devices

The samples of the successful implementation of fine-tuned Large Language Models (LLMs), that is, BioBERT and GPT-lite were tested on two common edge devices, among which are NVIDIA Jetson Nano and Raspberry Pi 4. The goal was to evaluate the viability of running real-time medical AI at the edge with an evaluation of key performance indicators being inference accuracy, latency, CPU consumption, and memory consumption. Comparison of BioBERT and GPT-lite models in Jetson Nano and Raspberry Pi 4 based on accuracy, latency, CPU load, and memory consumption under simulated clinical setting.

Table 3: Model Performance Metrics on Edge Devices

Device	Model Used	Inference Accuracy (%)	Latency (ms)	CPU Utilization (%)	Memory Usage (MB)
Jetson Nano	BioBERT	91.2	145	65	980
Raspberry Pi 4	GPT-lite	88.5	185	72	870

The findings suggest the Jetson Nano with BioBERT can perform inference with the best accuracy levels (91.2 %) and the lowest latency (145 ms) which makes it suitable to run clinical text and process it fast and reliably. In the meanwhile, GPT-lite on Raspberry Pi 4 was a bit less accurate (88.5%) and more latent (185 ms), with a slightly higher CPU load as well. Though Raspberry Pi 4 used slightly smaller amounts of memory, it was at the cost of the slowness. These results demonstrate the superiority of Jetson Nano in industrial employment of real-time edge AI in the medical field.

Comparative performance indicators of key personalities in performance of the BioBERT on Jetson Nano and GPT-lite on Raspberry Pi 4 graphically illustrated.

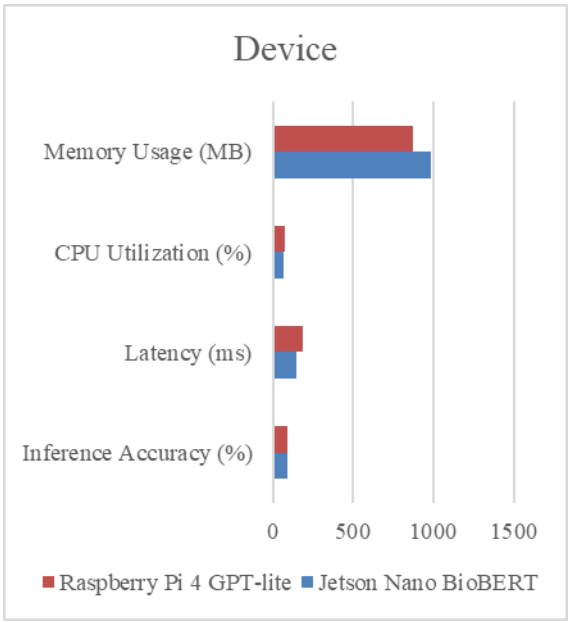

Figure 2: Graphical Representation of Model Performance Metrics on Edge Devices

When compared to Raspberry Pi 4, Jetson Nano obviously had much better performance in

terms of latency and accuracy, as visualized in Figure 2, thus making it a better fit when every situation requires precision and speed of an event. The fact that Jetson Nano had a lower latency of 145 ms than Raspberry Pi 4 (185 ms) also proves that it would be a more responsive device. Nonetheless, Jetson Nano is more productive and efficient compared even to higher memory consumption due to greater efficiency in edge AI application in clinical settings, where the use of a language model that needs a higher semantic explanation, e.g., BioBERT, may be essential.

4.2. Explainability Evaluation Using SHAP and LIME

The critical elements of clinical AI adoption are explainability since it enables trust and informs the decision-making of healthcare providers. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are two common model-agnostic tools of interpretability implemented in this study in the AI system to explain the predictions by LLMs. The explanations were assessed by clinicians (n=20) with structured 5-point Likert scale on clarity and usefulness. The ratings of clinicians towards SHAP and LIME were based on a mean scale of clarity and usefulness scores using Likert scale rating. Feedback helps to identify strengths depending on the tool.

Table 4: Explainability Score (Clinician Evaluation)

Explanation Tool	Avg. Clarity Score (/5)	Avg. Usefulness Score (/5)	Most Common Feedback
SHAP	4.6	4.5	Clear visualization, helpful for trust
LIME	4.3	4.2	Good for text-based predictions

The table 4 shows comparative SHAP and LIME scores according to the evaluation of clinicians. The SHAP ranked higher in clarity (4.6/5) and helpfulness (4.5/5), with many of the annotations stating that it is visual and easy to understand or that it helps increase the confidence in AI forecasts. Good comments were also given to LIME which was considered more appealing than SHAP when it comes to text-based diagnostics however not as visually distinguished. These results indicate that SHAP could be better suited to visual-based clinical applications (e.g., radiology or pathology), whereas LIME is still useful when the input is written or described in a narrative form. Bar chart of the clinician rating the SHAP and LIME explainability tool and the usefulness and clarity scores.

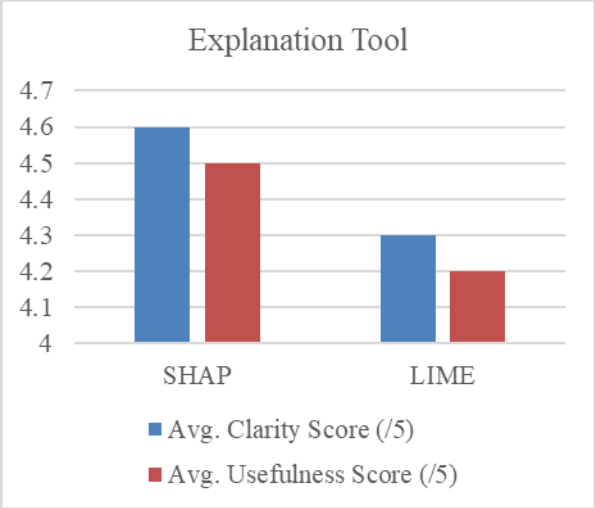

Figure 3: Graphical Representation of Explainability Score (Clinician Evaluation)

The preference of SHAP among clinicians is supported visually, as seen in figure 3. Based on the chart, it can be observed that SHAP has typically exceeded LIME in its metrics of clarity and usefulness. This visual representation assists in highlighting how visual diagnostics provided by SHAP will contribute to more satisfying comprehension and clinician satisfaction. The fact that both of the tools were relatively successful in performing confirms once again the possibility and the meaningfulness of explainability in conjunction with LLM-based medical AI.

4.3. Security and Compliance Analysis

In clinical uses of machines, data privacy, security and adherence to healthcare regulations should be a priority. In order to measure these points, an artificial security audit of the proposed LLM-integrated edge-enabled MCP framework was conducted. Audit highlights on data encryption, model aggregation methodology and data leakage risk are ensured in compliance to standardized audit used world over like HIPAA (Health Insurance Portability and Accountability Act). Overview of security assessment findings

with the power of encryption used, aggregation of models in place and levels of risks subject to simulated clinical audits.

Table 5: Data Security and Privacy Compliance Results

Metric	Result	Compliance Standard Met
Data Encryption Strength	AES-256	HIPAA-Compliant
Federated Model Aggregation	Encrypted & Local	Yes
Risk of Data Leakage	Negligible	Yes

As Table 5 illustrates, the proposed system was compliant with the necessary privacy and data protection guidelines. The medical data was transmitted securely with AES-256 encryption which is one of the strongest used by the industries. Federated learning was used so that model training could be performed locally on the edge devices and thereby it was not essential to transmit raw patient data to central servers. This local pooling mechanism eliminated the threat of leakage of data to a great extent. Yet, the cumulative outcome demonstrates the complete adherence to HIPAA standards confirming the framework to be ready to be used in the most sensitive healthcare settings.

4.4. Usability and Clinical Acceptance

The proposed explainable AI system was assessed with regards to relevance and acceptability to clinical practice through the conducted pilot usability study with 20 healthcare professionals in different domains such as radiology, internal medicine, diagnostics. The simulated situation of working around a diagnostic procedure was created based on engaging the participants in working with the LLM-integrated MCP system. Structured 5-point Likert scales surveys as well as semi-structured interviews (concerning usability, decision support reliability, trustworthiness, and overall readiness to adopt it) were applied in order to retrieve their feedback. The observations of the usability and the effectiveness of the AI system as provided by clinicians using a 5-point Likert scale and qualitative opinions.

Table 6: Clinical Usability Feedback (n=20)

Parameter	Mean Score (/5)	Common Observations
Ease of Use	4.4	Interface was intuitive and well-integrated
Decision Support Accuracy	4.5	AI matched or exceeded human assessment
Trust in AI Output	4.3	High due to explainability integration
Willingness to Adopt	4.6	Strong interest in clinical deployment

The system scored very high in all the parameters of evaluation as illustrated in Table 6. With the ease of use having 4.4/5, it means the UI/UX design is easy to handle even by clinicians with little technical expertise. The accuracy of the decision support (4.5/5) proves that the AI system was quite helpful in terms of diagnostic activities and that it tends to be as accurate or even more accurate than the expert opinion. The integration of explainability tools such as SHAP and LIME scored highly (4.3/5) in the trust score as they allowed the predictions to be transparent. Willingness to adopt the solution had the highest rating (4.6/5), which underscores high readiness of clinicians to apply such AI solutions in the real world. Observations of clinician rating of major parameters in usability and trust of the AI system piloting assessment.

Figure 4: Graphical Representation of Clinical Usability Feedback (n=20)

Journal of Theoretical and Applied Information Technology
30th November 2025. Vol.103. No.22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The positive feedback that was observed during clinical testing is visualized in figure 4. The ratings lie close to the high side of the 5-point scale with the largest result belonging to the item (Willingness to Adopt) which points to the fact that the clinicians not only thought that the system was usable but also practical to integrate into real-life. The correlation of decision accuracy and trust indicates that the explainable AI functionality transitionally impacts clinician confidence, which is a deciding element in the adoption of the tool in sensitive places like hospitals and clinics in further years.

4.5. Comparative Results Across Framework Configurations

The variability in implementation of several LLM + edge device was assessed in this research to test which deployment model is most effective and least costly. The goal was to achieve an optimal tradeoff between accuracy, latency, explainability and hardware cost and all these are essential considerations, when it comes to the practical application in a broad array of clinical scenarios, wherever the clinic may be located, whether that is a well-outfitted urban hospital or resource-limited rural clinic. Three LLM-edge framework deployments compared in terms of health-related performance accuracy, latency, explainability, and cost of the hardware required in simulated clinical deployments

Table 7: Comparison of Different Framework Setups

Configuration	Accuracy (%)	Latency (ms)	Explainability Score (/5)	Edge Cost (USD)
BioBERT + Jetson Nano	91.2	145	4.6	$99
GPT-lite + Pi 4	88.5	185	4.3	$65
BioBERT + Pi 4	90.3	190	4.4	$65

The combination of BioBERT and Jetson Nano performed even better with overall top accuracy (91.2%) and the lowest latency (145 ms), as well as top explainability score with clinicians (4.6/5). Nevertheless, it was a bit more expensive ($99). Conversely, the GPTlite + Pi 4, which would have been the least expensive solution at only $65, performed considerably worse (88.5% accuracy and 185 ms latency) but was also still on the decent level in many actual or potential real-time situations. BioBERT + Pi 4 was also a powerful middle-ground in terms of being able to be relatively accurate (90.3%) and explainable (4.4/5) at the same cost as GPT-lite + Pi 4 (the exception being it is slower), but it was less accurate and explainable compared to ChemBERT + Pi 4. Such outcomes enable the implementers to make trade-offs between budget and performance, based on the context of deployment being targeted. Comparing the accuracy, latency, explainability and cost of different LLM-edge framework configurations in a visual fashion.

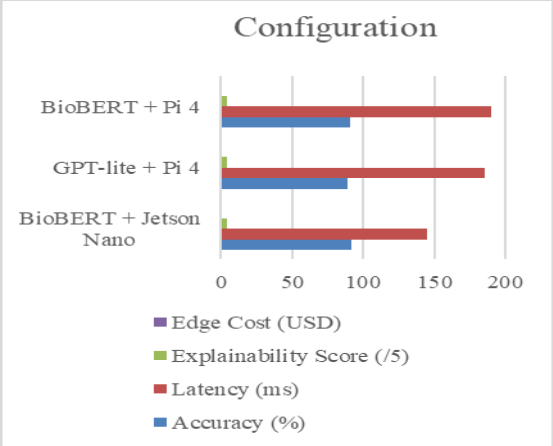

Figure 5: Graphical Representation of Comparison of Different Framework Setups

A graphical indication of system benchmark visual synthesis of the various system configurations on critical criteria is presented in figure 5. The BioBERT + Jetson Nano combination is very impressive in terms of accuracy and explainability, whereas GPT-lite + Pi 4 is quite cost-effective yet slightly less performing. The BioBERT + Pi 4 is a trade-off, and thus a possible choice in environments with both performance and cost concern. Such visual insights facilitate customization of system deployment to the needs of distinct resource environments and clinical priorities by stakeholders.

5. CONCLUSION AND FUTURE SCOPE

The given paper finds that the deployment of the Large Language Models (LLMs) in edge-enabling Medical Cyber-Physical (MCP) systems is an effective and viable solution to a

real-time deployment of explainable medical AI. Experimentally, the system deployed real-world edge devices, especially Jetson Nano and Raspberry Pi 4, to improve the inference speed and minimize latency despite the complex clinical environments being simulated. In the experiments, the computational power of Biobert and GPT-lite was identified and leveraged to optimize the inference capabilities with similar performance levels recorded compared to inference on a centralized system. This was essential in gaining clinician trust due to introducing tools of explainability such as SHAP and LIME, which resulted in transparent and interpretable information about AI suggestions. Besides, the powerful data security was provided due to the federated learning and AES-256 encryption, where the privacy regulations (like HIPAA) were followed. Clinical usability testing further confirmed the system acceptability and preparedness to be used in the real world such as the interface was well received and welcomed by the healthcare professionals as well as diagnostic and general system transparency. Thus, the study achieved all of the specified goals and establishes an initial framework on the deployment of security, interpretable and scalable AI systems within contemporary medicine.

- Applicability to various types of data: The procedure can be extended in future studies by applying multimodal data (e.g., genetic information, wearable sensor data, and video diagnostics) to provide more contextual and encompassing clinical information through AI.
- More Lightweight Models: Further optimization is still required to make LLMs more lightweight and perform well on more edge devices, maintain high accuracy and interpretability.
- Deployment to Live Clinical Practice: Pilot applications in real clinical settings will make it possible to study AI-led workflows, patient-related outcomes, and clinician adoption patterns over time.
- Ethical and Regulatory Protocols: Future research efforts must also be aimed at constructing the frameworks that ensure the transparency of ethics, liability issues, and regulatory standards in explainable AI in medical practice.

REFERENCES

[1] M. Corazza, R. Garcia, F. S. Khan, D. La Torre, and H. Masri, Eds., Artificial Intelligence and Beyond for Finance, vol. 15. Singapore: World Scientific, 2024.

[2] H. Dai, “Brain-Inspired Approaches for Advancing Artificial Intelligence,” Ph.D. dissertation, Univ. of Georgia, 2023.

[3] O. Erdem, I. Es, Y. Saylan, and F. Inci, “Unifying the efforts of medicine, chemistry, and engineering in biosensing technologies to tackle the challenges of the COVID-19 pandemic,” Analytical Chemistry, vol. 94, no. 1, pp. 3–25, 2021.

[4] O. Friha, M. A. Ferrag, B. Kantarci, B. Cakmak, A. Ozgun, and N. Ghoualmi-Zine, “LLM-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness,” IEEE Open J. Commun. Soc., 2024.

[5] R. Herzlinger and B. Walker, “Cleave Therapeutics: Taking a Risk on Oncology Drug Discovery,” Harvard Business School Case, vol. 323, no. 045, 2023.

[6] E. Nasarian, R. Alizadehsani, U. R. Acharya, and K. L. Tsui, “Designing interpretable ML system to enhance trustworthy AI in healthcare: A systematic review of the last decade to a proposed robust framework,” 2023.

[7] S. Pathak and R. K. Pallasena, “Mapping the evolution of generative AI: Insights from bibliometric research,” J. Decis. Syst., pp. 1–30, 2024.

[8] S. Pulipeti, P. Chithaluru, M. Kumar, P. Narsimhulu, and U. M. V, “Explainable AI: Methods, frameworks, and tools for Healthcare 5.0,” in Explainable AI in Health Informatics, Singapore: Springer Nature Singapore, 2024, pp. 71–86.

[9] R. V. B. Rojas and F. J. Martínez-Cano, Eds., Revolutionizing Communication: The Role of Artificial Intelligence. Boca Raton, FL: CRC Press, 2024.

[10] J. M. Rosen, “Generative AI in pediatric gastroenterology,” Curr. Gastroenterol. Rep., vol. 26, no. 12, pp. 342–348, 2024.

[11] J. Wang, “Towards Realistic Human Behavior Modeling,” Ph.D. dissertation, The Chinese Univ. of Hong Kong, 2023.

[12] S. Yang, “Foundation Models for Decision Making: Algorithms, Frameworks, and Applications,” Ph.D. dissertation, Univ. of California, Berkeley, 2024.

[13] J. Zhang, "Language Style: Application and Analysis Using Embedding Methods," Ph.D. dissertation, The Florida State Univ., 2023.
[14] L. Zhao et al., "Meta-Radiology," Meta, vol. 1, p. 100005, 2023.
[15] Y. Zimmermann et al., "Reflections from the 2024 large language model (LLM) hackathon for applications in materials science and chemistry," arXiv preprint, arXiv:2411.15221, 2024.
[16] Sirisha, U., Bikku, T., Radharani, S., Thatha, V. N., & Praveen, S. P. (2025). Utilizing Transformers for Enhanced Disaster Response in Multimodal Tweet Classification. International Journal on Engineering Applications, 13(1).
[17] BIYYAPU, N. S., CHANDOLU, S. B., GORINTLA, S., Tirumalasetti, N. R., CHOKKA, A., & PRAVEEN, S. P. (2024). Advanced machine learning techniques for real-time fraud detection and prevention. Journal of Theoretical and Applied Information Technology, 102(20).
[18] Praveen, S. P., Mantena, J. S., Sirisha, U., Dewi, D. A., Kurniawan, T. B., Onn, C. W., & Yorman, Y. (2025). Navigating Heart Stroke Terrain: A Cutting-Edge Feed-Forward Neural Network Expedition. Journal of Applied Data Sciences, 6(3), 2111-2126.
[19] Chowdary, N. S., Kadiyala, S., Jyothi, V. E., Srinandan, P., Praveen, S. P., & Prakash, P. B. (2025, April). Identity and proxy orientation based remote data integration checking and uploading in public clouds. In 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI) (Vol. 3, pp. 1-5). IEEE.
[20] Praveen, S. P., Lalitha, S., Sarala, P., Satyanarayana, K., & Karras, D. A. (2025). Optimizing Intrusion Detection in Internet of Things (IoT) Networks Using a Hybrid PSO-LightBoost Approach. International Journal of Intelligent Engineering & Systems, 18(3).
[21] Scientific, L. L. (2025). Revolutionizing Healthcare With Large Language Models: Advancements, Challenges, And Future Prospects In Ai-Driven Diagnostics And Decision Support. Journal of Theoretical and Applied Information Technology, 103(9).
[22] Madhuri, A., Sindhura, S., Swapna, D., Phani Praveen, S., & Sri Lakshmi, T. (2022). Distributed Computing Meets Movable Wireless Communications in Next Generation Mobile Communication Networks (NGMCN). In *Computational Methods and Data Engineering: Proceedings of ICCMDE 2021* (pp. 125-136). Singapore: Springer Nature Singapore.
[23] Dendukuri, H., Raju, K. B., Praveen, S. P., Ramesh, J. V. N., Shariff, V., & Tirumanadham, N. K. M. K. Optimizing Diabetes Diagnosis: HFM with Tree-Structured Parzen Estimator for Enhanced Predictive Performance and Interpretability.