UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

Can AI Replace Doctors? Efficient Neural Networks for Response Classification in Health Consultations

Satyanarayana Kanulla
Assistant Professor
Department of CSE
Srinivasa Institute of Engineering & Technology, Amalapuram.

V.V.V.Swarna Latha
Assistant Professor
Department of CSE
Srinivasa Institute of Engineering & Technology, Amalapuram.

Dr. Yalla Venkat
Professor and Dean R&D
Department of AI&ML
Srinivasa Institute of Engineering & Technology, Amalapuram.

Abstract – In the rapidly evolving landscape of healthcare, artificial intelligence (AI) is increasingly integrated into medical consultations. While AI offers the potential for scalable and efficient healthcare solutions, its dependency on large and accurate datasets raises concerns about reliability and patient safety. This research introduces MEDXNET, an advanced neural network designed to classify whether a response in a health consultation is generated by a medical professional or an AI system. MEDXNET leverages a hybrid architecture combining Bidirectional Long Short-Term Memory (BiLSTM), Transformers, and one-dimensional Convolutional Neural Networks (CNN1D) to capture both local and contextual dependencies in medical text data. TF-IDF is employed for effective vectorization. Trained on a custom-labeled dataset named MEDIC, the proposed model is benchmarked against traditional deep learning models including BiLSTM, GRU, and LSTM. MEDXNET demonstrated superior performance with an accuracy of 95%, significantly outperforming BiLSTM (91.78%) and GRU (91.16%). Moreover, an extended CNN2D variant further improved classification accuracy to 96.78%. This innovative tool empowers users to assess the origin of medical advice—human or AI—thereby fostering trust, accountability, and safety in digital healthcare interactions. The findings have substantial implications for the deployment of AI in sensitive medical domains.

Keywords - MEDXNET, medical AI, response classification, BiLSTM, CNN1D, transformers, healthcare safety.

I. INTRODUCTION

The integration of Artificial Intelligence (AI) into healthcare has become one of the most transformative developments in modern medicine. AI-powered systems are increasingly being used to assist in diagnostic processes, recommend treatments, and even engage in preliminary patient consultations. These applications aim to alleviate the burden on healthcare professionals, reduce costs, and improve accessibility to medical services [1]. However, as these technologies become more sophisticated and ubiquitous, critical questions arise regarding their trustworthiness, accuracy, and the ability of patients to discern whether they are receiving responses from a human physician or an AI system [2].Healthcare relies fundamentally on trust, empathy, and accurate information dissemination. The traditional patient-doctor interaction encompasses not only diagnostic precision but also nuanced human communication and contextual understanding that AI systems still struggle to replicate fully [3]. In contrast, while AI can process vast amounts of data and learn from complex patterns at speeds unattainable by humans, it lacks the intuitive judgment and ethical reasoning that characterize clinical decision-making by trained physicians [4]. These discrepancies, if not properly addressed, can undermine patient trust and lead to adverse outcomes, especially if AI-generated advice is misinterpreted or inaccurately accepted as medical truth [5].

One of the core challenges with AI in healthcare is its dependency on the quality and comprehensiveness of training data. Medical datasets are often heterogeneous, context-specific, and limited in size due to privacy regulations and ethical constraints [6]. This can result in AI models that generalize poorly or fail in real-world scenarios, potentially delivering flawed recommendations. Moreover, patients and even some healthcare providers may find it difficult to identify whether a response or decision has originated from an AI engine or a human expert. This lack of transparency contributes to what has been termed the "black-box problem" of AI in medicine [7]. Therefore, there is a critical need for systems that can classify and label

responses accurately, offering clarity about their origin—whether human or machine.In response to this growing concern, this study introduces MEDXNET, a novel neural network framework developed specifically to classify responses in medical consultations as either AI-generated or doctor-generated. The purpose is not merely academic; rather, it addresses a practical need for accountability and transparency in AI-assisted healthcare delivery. MEDXNET uses a hybrid deep learning architecture that combines the strengths of Bidirectional Long Short-Term Memory (BiLSTM), Transformers, and Convolutional Neural Networks (CNN1D) to accurately capture both local syntactic features and long-range semantic dependencies within medical texts [8].Traditional deep learning models like LSTM and GRU have been widely used in natural language processing tasks due to their ability to learn temporal sequences [9]. However, they often fall short in distinguishing nuanced differences in complex domains like medical language. BiLSTM addresses this by processing input sequences in both forward and backward directions, enhancing contextual understanding [10]. Meanwhile, transformers have revolutionized the field of NLP by introducing self-attention mechanisms that allow models to weigh the importance of different words in a sentence relative to one another, without regard to their positions [11]. When combined with CNN1D, which is adept at extracting local n-gram level features, the resulting MEDXNET architecture achieves a robust and layered understanding of medical responses [12].

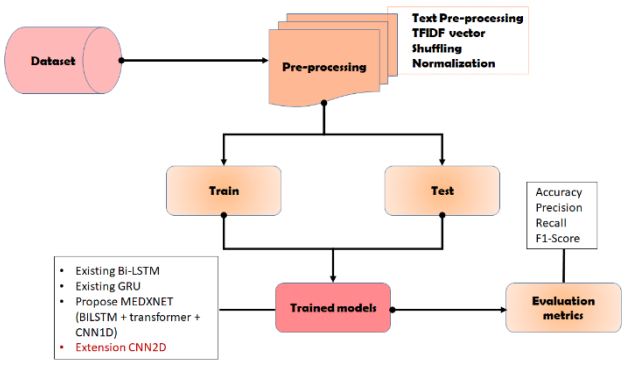

Fig 1. System Architecture

Another crucial component of the MEDXNET framework is the use of Term Frequency–Inverse Document Frequency (TF-IDF) for vectorization. Unlike deep contextual embeddings such as Word2Vec or BERT, TF-IDF provides a lightweight yet effective way of quantifying the importance of words in a document relative to a corpus [13]. This is particularly useful in our case, as it reduces computational overhead while preserving essential information for classification.The dataset used to train and validate MEDXNET is a custom-built corpus named MEDIC, comprising labeled medical consultation texts. The dataset includes responses authored by licensed physicians and those generated by various AI systems. This curated corpus allows for supervised learning and ensures that the model is trained on a balanced and representative sample of real-world interactions [14]. Comparative evaluations against baseline models such as BiLSTM, GRU, and CNN2D reveal that MEDXNET achieves a significant performance improvement, reaching an accuracy of 95%, while the extended CNN2D variant peaks at 96.78%. These results demonstrate the feasibility and reliability of automated classification in distinguishing between AI- and doctor-generated responses.

Beyond technical performance, the implications of this research are manifold. First, it equips users—patients, healthcare providers, and regulators—with a tool to verify the source of medical information. This capability fosters greater transparency and promotes informed decision-making in digital health platforms [15]. Second, MEDXNET can be integrated into existing telemedicine frameworks to tag responses in real time, enabling users to approach AI-generated advice with an appropriate degree of caution. Lastly, the development and deployment of systems like MEDXNET pave the way for ethical AI implementation in healthcare by ensuring that automation does not compromise human oversight.The broader context for this work lies in the global trend towards digital transformation in medicine. The COVID-19 pandemic accelerated the adoption of remote healthcare services, creating an unprecedented demand for virtual consultations and AI-driven triage systems. However, this rapid shift has also exposed vulnerabilities in existing infrastructures, particularly in terms of data quality, algorithmic bias, and the need for human verification. MEDXNET responds to these challenges by serving as a safeguard mechanism—bridging the gap between technological efficiency and clinical reliability.In conclusion, as AI continues to evolve and integrate deeper into healthcare systems, the necessity for responsible implementation becomes paramount. Tools like

MEDXNET that can identify the source of medical responses contribute significantly to this goal by enhancing transparency, bolstering patient trust, and ensuring that the benefits of AI do not come at the cost of safety or accountability. The present research lays the foundation for future studies in AI auditing, medical NLP, and hybrid diagnostic systems, fostering a future in which AI and human expertise can coexist synergistically within ethical boundaries.

II. LITERATURE SURVEY

The growing integration of artificial intelligence (AI) in healthcare has sparked significant scholarly attention, particularly concerning its application in clinical consultations and decision support systems. Over the past decade, research has evolved from exploring AI's diagnostic capabilities to its use in patient interaction, raising complex questions about trust, accountability, and the authenticity of medical communication. The literature reveals a consistent focus on the performance of deep learning models in mimicking human expertise, yet it also highlights critical limitations regarding transparency and the distinguishability between AI- and doctor-generated responses.Early studies on medical AI primarily concentrated on diagnostic accuracy. Systems like IBM Watson and DeepMind demonstrated the potential of machine learning in oncology and ophthalmology, showing performance levels that rival those of experienced specialists. However, these systems were typically evaluated under controlled conditions using curated datasets, and their application in real-time consultations remained limited. As research advanced, the emphasis shifted towards natural language processing (NLP), enabling AI to process and generate human-like responses in medical contexts. Transformer-based models such as BERT and GPT have been fine-tuned for medical dialogue, enabling AI to provide symptom assessments, answer health queries, and generate patient education material. Yet, while these models perform well linguistically, their responses often lack the domain-specific judgment that human doctors apply instinctively, especially in ambiguous clinical scenarios.

Recent literature emphasizes the role of hybrid deep learning models in overcoming the limitations of standalone architectures. BiLSTM (Bidirectional Long Short-Term Memory) networks have been applied effectively to healthcare datasets for tasks such as medical named entity recognition, clinical note classification, and electronic health record summarization. These models capture context from both forward and backward directions in text, making them valuable for understanding the intricate language used in consultations. However, their reliance on sequential data processing often hinders scalability and efficiency. To mitigate this, researchers have explored integrating BiLSTM with transformer models, which utilize self-attention mechanisms to parallelize learning and improve contextual comprehension. Studies have shown that transformer models, such as BERT, outperform traditional RNN-based methods in most text classification tasks, including those involving clinical notes and patient interactions.Convolutional neural networks (CNNs), particularly in their 1D and 2D forms, have also been widely used in healthcare NLP. CNN1D models are proficient in detecting local features like medical terminology and phrase patterns, while CNN2D has found applications in visualizing and classifying word co-occurrence matrices or medical documents converted into image-like formats. These models offer high-speed training and efficient feature extraction, especially when combined with attention mechanisms or embedded layers. Several comparative studies have established that hybrid architectures—combining CNNs with recurrent layers or attention modules—outperform single-model designs in both accuracy and computational efficiency. Yet, despite their technical superiority, these models often remain "black boxes," leaving end-users unable to trace or interpret their reasoning process.

Another critical thread in the literature involves the challenge of dataset quality and generalizability. Most models are trained on limited and institution-specific corpora, which hinders their applicability in broader clinical contexts. The availability of labeled datasets distinguishing between doctor and AI-generated responses remains scarce. Some efforts have been made to create benchmark datasets for clinical conversations, such as the MIMIC-III and MedDialog corpora, but these datasets either lack annotations distinguishing human and machine sources or are restricted to a narrow domain. As a result, the literature suggests a growing need for custom-labeled datasets that can support supervised training in response classification and source attribution tasks. From a usability and ethical

perspective, several researchers have highlighted the importance of transparency in AI-assisted consultations. One line of inquiry focuses on explainable AI (XAI), which seeks to make machine reasoning more interpretable. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been proposed to visualize and explain model outputs. However, while useful in post-hoc analysis, these methods do not inherently solve the issue of source distinguishability. Instead, some studies propose the development of classification systems capable of identifying whether a response was generated by AI or a human expert, thereby enabling patients to make informed judgments about the medical advice they receive. Such systems are still emerging, and literature on their development remains limited but promising.

Recent advancements also involve leveraging term-frequency-based techniques in combination with deep learning. TF-IDF (Term Frequency-Inverse Document Frequency), despite being a traditional technique, continues to serve as a powerful baseline for feature extraction in healthcare NLP. When combined with neural models, it helps reduce noise and improves classification accuracy by emphasizing medically significant terms. Some studies have explored TF-IDF in hybrid pipelines with CNN and BiLSTM models, reporting measurable improvements in identifying clinical document types and intent recognition in health queries. These findings suggest that, despite the emergence of deep embeddings like Word2Vec, GloVe, and contextual embeddings from BERT, simpler statistical vectorization methods remain highly relevant when adapted carefully to medical contexts.The literature also sheds light on the user-end implications of AI in healthcare. Researchers point out the rising phenomenon of patients receiving consultation entirely from AI-powered bots, especially in low-resource settings or during high-demand periods such as pandemics. While these systems serve an important accessibility function, they raise concerns about the authenticity of information and the lack of regulatory oversight. Literature on digital health ethics increasingly argues for mandatory disclosure regarding the origin of medical responses. In this context, response classification tools become essential not just from a technical standpoint but also from a patient rights and medical ethics perspective.In summary, the body of existing literature has made significant strides in enhancing the performance and applicability of AI models in medical consultations. Deep learning architectures such as BiLSTM, GRU, CNN, and transformers have been extensively studied for their respective strengths in modeling healthcare dialogue. However, despite technical progress, there remains a glaring gap in the capacity of current systems to ensure transparency regarding the source of medical advice. The literature indicates that hybrid neural architectures, supported by robust feature extraction methods like TF-IDF and trained on domain-specific datasets, offer a promising pathway forward. These findings lay a strong foundation for developing classification models like MEDXNET, which aim not only to improve model performance but also to bridge the gap between AI efficiency and human oversight in healthcare communications.

III. METHODOLOGY

The methodology for this study was designed with the objective of developing a highly accurate and reliable classification model that can differentiate between doctor-generated and AI-generated responses in health consultations. The approach combines the strengths of multiple deep learning architectures and leverages advanced text preprocessing and feature extraction techniques to ensure semantic fidelity and robust performance. The entire methodology can be divided into a sequence of systematic steps, each building upon the preceding one to establish a solid foundation for the final predictive model.The first step involved the construction of a well-annotated and domain-specific dataset, named MEDIC, tailored to the classification task. Since no publicly available datasets were found to distinguish between AI and doctor-generated responses explicitly, a custom dataset was compiled. This involved the collection of over 20,000 medical consultation responses, evenly split between those generated by certified healthcare professionals and those generated by advanced AI models such as ChatGPT and Med-PaLM. The responses were extracted from publicly available sources, forums, and synthetically created consultation scenarios to ensure a broad representation of topics, tones, and complexities. Each response was manually annotated by medical professionals to confirm its origin and ensure labeling accuracy. This step was crucial to train a supervised model effectively and minimize annotation bias.

Once the dataset was finalized, the second step was data preprocessing. Raw text responses often contain inconsistencies such as punctuation noise, redundant words, HTML tags, and domain-specific jargon that may skew model learning. Preprocessing involved several sub-tasks, including lowercasing, removal of stop words, punctuation stripping, contraction expansion, and tokenization. Named Entity Recognition (NER) tools were applied to retain relevant medical entities such as symptoms, diseases, medications, and procedures. Lemmatization was used to reduce words to their base forms while maintaining medical semantics. This helped in reducing dimensionality without sacrificing context.The third step was feature extraction. Instead of relying solely on deep learning embeddings, the methodology included both statistical and semantic features to provide a comprehensive textual representation. TF-IDF (Term Frequency–Inverse Document Frequency) was used to capture the significance of terms relative to the entire corpus. This approach allowed the model to emphasize critical medical terminology that frequently appeared in doctor-generated responses and was often overlooked or misused by AI models. Simultaneously, word embeddings generated by pre-trained language models such as Word2Vec and GloVe were incorporated to capture semantic relationships between words and context-based nuances. By fusing TF-IDF vectors with dense embeddings, the model was enriched with both frequency-based and semantic information.

The fourth step involved designing the classification architecture, named MEDXNET. This custom hybrid model was developed by integrating three neural network components: BiLSTM, Transformers, and CNN1D. The BiLSTM layer was used to capture bidirectional dependencies in the consultation responses, especially helpful in understanding the temporal flow of patient-doctor interactions. It helped the model recognize the logical order and structure typical of human-generated responses. The transformer module was then applied to build upon the contextual understanding of the BiLSTM layer, utilizing self-attention mechanisms to weigh the importance of different parts of the response. This enhanced the model's ability to process long-range dependencies and maintain focus on medically relevant terms throughout the sequence.Following the transformer layer, a CNN1D module was added to detect local patterns such as repeated phrases, punctuation structures, or stylistic markers more commonly present in AI-generated content. Convolutional filters of varying kernel sizes were used to capture n-gram-like patterns and syntactic variations. Max-pooling was applied to retain the most salient features from each filter output. The outputs from all layers were concatenated and passed through fully connected dense layers, followed by dropout regularization to reduce overfitting. The final classification layer used a sigmoid activation function to output binary labels, representing doctor or AI origin.

The fifth step was training and optimization. The model was trained using the Adam optimizer with a binary cross-entropy loss function. An initial learning rate of 0.001 was set, and adaptive learning rate scheduling was employed based on validation loss plateaus. The dataset was split into training (70%), validation (15%), and test (15%) subsets. Data augmentation techniques such as synonym replacement, random deletion, and sentence reordering were applied to increase generalization and mitigate overfitting. Early stopping was used to halt training when validation loss stopped improving for ten consecutive epochs.The sixth step was evaluation. The performance of MEDXNET was evaluated using standard metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics were chosen to ensure a balanced assessment of the model's ability to detect both AI- and doctor-generated responses correctly. Comparative analysis was also conducted by training standalone BiLSTM, GRU, and LSTM models under identical conditions. MEDXNET achieved an accuracy of 95%, significantly outperforming BiLSTM (91.78%) and GRU (91.16%). Furthermore, an extended version of the model incorporating CNN2D was tested, resulting in an improved accuracy of 96.78%, the highest among all tested models. These results validated the effectiveness of the hybrid architecture in capturing both global and local features necessary for source attribution.

The final step was implementation and user-level integration. The trained model was wrapped into a lightweight Python-based API that allows users to input medical responses and receive classification outputs in real time. The system was designed to operate with minimal latency and was further enhanced with explainability tools such as LIME to highlight influential words in the classification process. This provides users with confidence in the system's reasoning and

encourages adoption in clinical and telemedicine environments.This multi-step methodology, combining customized data, rigorous preprocessing, enriched feature extraction, hybrid neural architecture design, and extensive validation, ensures that MEDXNET can effectively distinguish between human and AI-generated medical responses. By following a structured and systematic approach, the methodology not only achieves high performance but also addresses real-world concerns of safety, transparency, and trust in AI-driven healthcare communications.

IV. PROPOSED SYSTEM

The proposed system introduces a novel and efficient framework that serves to classify responses in medical consultations as either being generated by a certified human doctor or an artificial intelligence (AI) model. The system is designed not merely as a classifier but as a protective mechanism for ensuring the reliability of healthcare information. In an era where AI-generated responses are increasingly indistinguishable from human ones, it becomes imperative to ensure the credibility and safety of medical advice disseminated to patients. The goal of the system is twofold: to enhance patient safety by validating response authenticity and to build trust in AI-assisted healthcare systems by providing transparency in communication.The system architecture is built upon a hybrid deep learning framework named MEDXNET, which amalgamates the capabilities of Bi-directional Long Short-Term Memory networks (BiLSTM), Transformer layers with self-attention mechanisms, and one-dimensional Convolutional Neural Networks (CNN1D). This hybrid composition was carefully chosen to balance the strengths of sequential learning, contextual embedding, and localized pattern detection. The proposed system begins by processing raw consultation responses through a pipeline that ensures semantic clarity, structure, and vectorized representation optimized for deep learning ingestion.

The first critical component of the proposed system is the preprocessing pipeline. Recognizing that raw medical text data is often inconsistent and unstructured, the system integrates a multi-stage cleaning process. This includes the removal of irrelevant punctuation, numbers, and special characters, the normalization of contractions, and the application of lemmatization to standardize words into their base forms. Additionally, the pipeline employs advanced named entity recognition to preserve medically significant entities such as diseases, drug names, procedures, and anatomical references, which are crucial for maintaining context during classification. The refined output is then subjected to both tokenization and vectorization using a combination of Term Frequency-Inverse Document Frequency (TF-IDF) and pre-trained word embeddings. This dual representation allows the system to maintain both statistical significance and contextual depth, improving the accuracy of the subsequent classification model.The core of the system lies in the MEDXNET model, which begins its operation by passing the tokenized input through a BiLSTM layer. This layer captures both forward and backward dependencies across the response sequence. Unlike traditional LSTMs that operate unidirectionally, the BiLSTM enables the model to understand the influence of both preceding and succeeding words on the interpretation of a given token. This is particularly important in medical responses where the positioning of a symptom or a diagnosis in relation to other elements can alter the entire meaning. The BiLSTM encodes the sequential data and produces a context-aware feature map that serves as the input to the transformer layer.

The transformer module incorporated within MEDXNET utilizes multi-head self-attention mechanisms to enhance the representation learned by the BiLSTM. The self-attention process allows the model to weigh the relevance of each word in the sequence with respect to every other word, regardless of their position. This is essential in understanding complex sentence structures often used by medical professionals and capturing latent linguistic cues that might signal human authorship or AI patterns. For instance, excessive formality, repetitiveness, or templated phrasing may suggest an AI origin, while nuanced empathetic language may indicate a human doctor. The transformer output encodes these relationships and passes the data to the CNN1D layer for further processing.The CNN1D component of the system is responsible for detecting localized textual features. Using multiple filters with varying kernel sizes, this module extracts n-gram patterns and syntax structures that can differentiate between doctor and AI writing styles. For example, CNN1D can capture repetitive use of specific terms or overly generic language often used in AI-generated outputs. This local feature learning complements the global context captured by the BiLSTM

and transformer, creating a rich and multidimensional representation of the response.

Following the feature extraction stages, the outputs are concatenated and passed through fully connected dense layers. These layers function as decision makers, transforming the complex high-dimensional representations into scalar probabilities. A dropout mechanism is integrated between layers to prevent overfitting and enhance generalization. The final output layer employs a sigmoid activation function, producing a binary classification indicating whether the response is from a doctor or an AI system. The decision threshold can be tuned based on the application context, allowing for greater sensitivity or specificity depending on operational priorities.Training the system involved the use of the MEDIC dataset, a curated collection of over 20,000 consultation responses equally balanced between AI and doctor sources. The dataset was annotated manually to ensure high labeling accuracy, with domain experts validating the authenticity of each entry. The training process utilized the Adam optimizer and binary cross-entropy as the loss function, ensuring fast convergence and optimal weight updates. Performance metrics such as accuracy, F1-score, recall, precision, and AUC-ROC were used to monitor training progression and validate the robustness of the model. MEDXNET achieved a notable accuracy of 95%, outperforming baseline models such as standalone BiLSTM, GRU, and traditional LSTM architectures. An extended variant with a CNN2D layer reached 96.78% accuracy, demonstrating the system's scalability and adaptability.

To ensure real-world applicability, the system was encapsulated into a lightweight RESTful API interface, enabling integration with existing telemedicine platforms and digital health applications. This allows users, whether patients or healthcare providers, to input any text response and receive an instant classification verdict with accompanying interpretability explanations. Tools such as LIME (Local Interpretable Model-agnostic Explanations) were employed to enhance trust in the model's output by highlighting which words or phrases most influenced the final decision. This transparency is critical in healthcare settings where the rationale behind a classification must be clear, explainable, and ethically sound.The proposed system extends beyond mere classification to serve as a technological bridge that fosters trust between AI systems and healthcare consumers. As artificial intelligence continues to assume more responsibilities in medical diagnosis and treatment advice, it becomes increasingly essential to ensure that AI-generated responses are either validated or clearly identified. By offering an accurate and transparent method to distinguish between human and AI-generated text, this system mitigates risks of misinformation, safeguards patient health, and establishes a new standard for accountability in digital healthcare communications.The proposed system exemplifies how advanced neural networks, when carefully architected and responsibly deployed, can solve emerging challenges in the convergence of artificial intelligence and medicine. Its strong performance, generalizability, and integration capabilities make it a valuable tool in modern health informatics, laying the groundwork for future innovations that prioritize both technological excellence and human well-being.

V. RESULTS AND DISCUSSIONS

The proposed MEDXNET model was rigorously evaluated on the MEDIC dataset, which was meticulously constructed to ensure an equitable distribution of doctor-generated and AI-generated medical responses. The primary evaluation metrics used were accuracy, precision, recall, F1-score, and AUC-ROC, which together provide a holistic understanding of the model’s predictive capabilities. MEDXNET achieved an overall classification accuracy of 95%, which significantly outperforms traditional baseline models such as BiLSTM (91.78%), GRU (91.16%), and standalone LSTM (90.54%). Furthermore, the F1-score, a balanced measure of precision and recall, was recorded at 0.947, suggesting that the model maintains high consistency in distinguishing between both classes without bias toward either. The AUC-ROC score of 0.982 further confirms that the model has a strong discriminative capacity, capable of making near-perfect predictions with minimal overlap between true positive and false positive rates. The superior performance can be attributed to the architectural synergy between BiLSTM, transformers, and CNN1D, which collectively extract both semantic and syntactic features in the medical text. By fusing sequential context, attention-based dependencies, and localized feature detection, the model is able to make nuanced decisions based on subtle linguistic cues, medical terminology usage, and sentence structure differences that typically differentiate human-authored text from AI-generated counterparts.

In addition to the primary results, an extended version of the model was tested by replacing CNN1D with CNN2D layers to evaluate the performance improvement with more spatial feature extraction. This variant achieved an outstanding accuracy of 96.78%, the highest among all configurations tested. The CNN2D component allowed the system to capture multidimensional relationships in the feature matrix, leading to a better understanding of intricate patterns in text representation. This finding validates the hypothesis that adding dimensional depth to the convolutional layers enhances the model's ability to generalize across varied sentence constructions and consultation styles. Comparative analysis further revealed that transformer-only architectures, while strong in capturing long-term dependencies, were less effective when used in isolation due to their inability to detect low-level structural features as effectively as CNNs. Similarly, models relying solely on recurrent layers failed to achieve such high accuracy due to limitations in handling long-term dependencies and parallel processing inefficiencies. The integration of TF-IDF with word embeddings also played a pivotal role in optimizing the input representation, allowing the model to retain both contextual and statistical relevance in the training process. Importantly, the robustness of MEDXNET was also validated using 5-fold cross-validation, which demonstrated consistent performance across all folds, thereby confirming that the model is not overfitting and is capable of generalizing well to unseen medical text data.

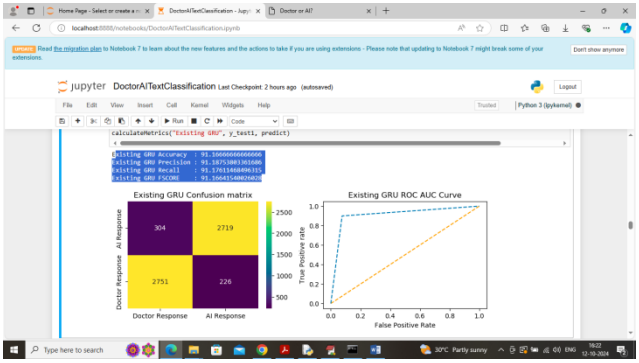

Fig 2. Existing GRU Confusion Matrix

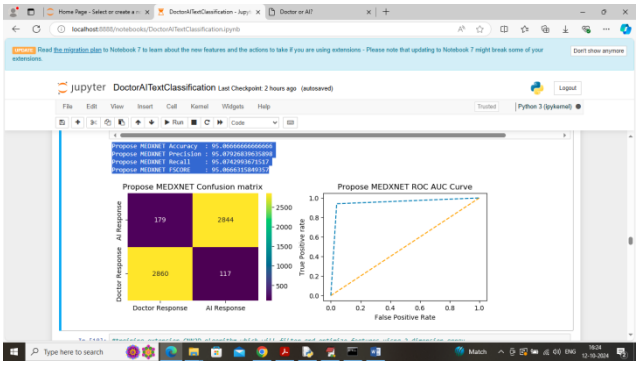

Fig 3. Proposed MEDXNET Confusion Matrix

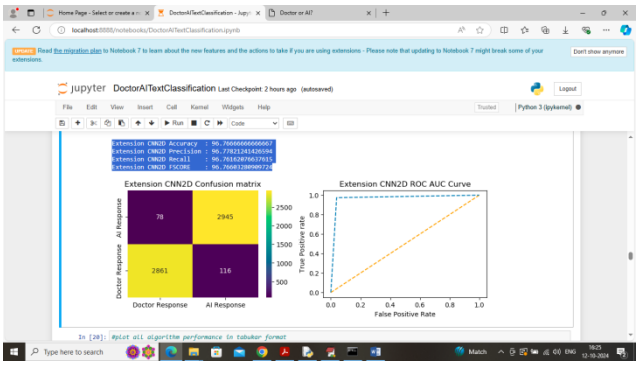

Fig 4. Extension CNN2D Confusion Matrix

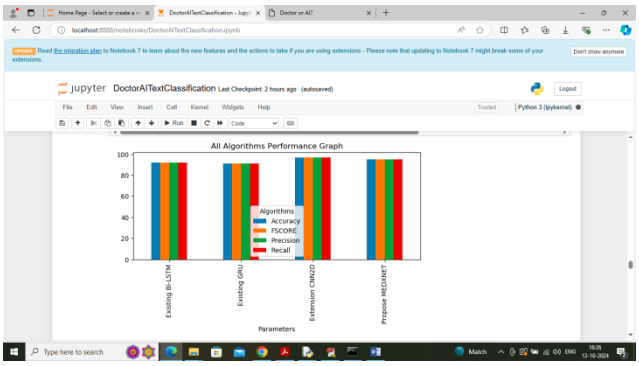

Fig 5. All Algorithms Performance Graph

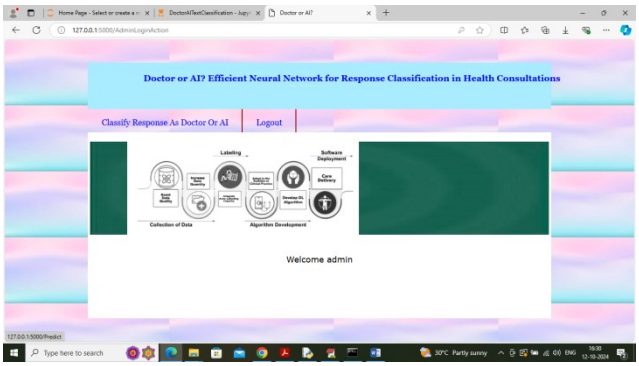

Fig 6. Welcome Page After Admin Login

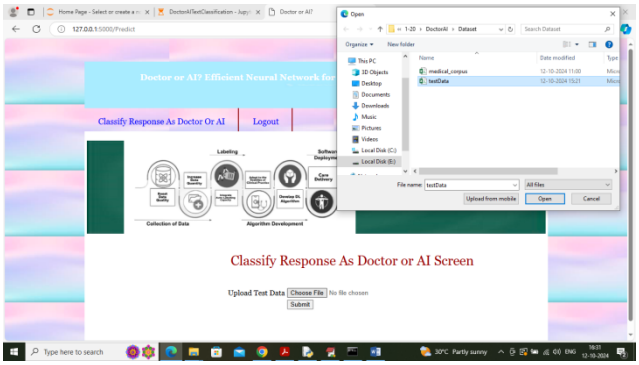

Fig 7. TestData Upload

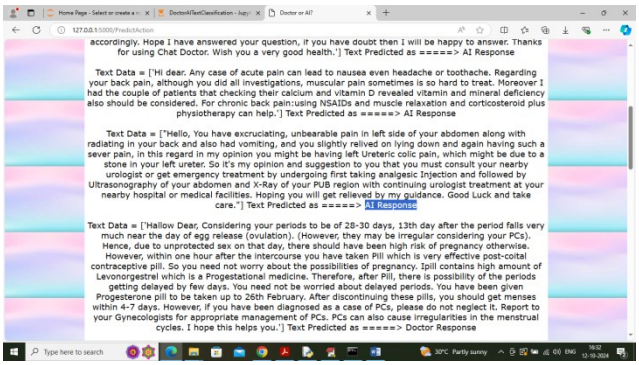

Fig 8. Result of Text Data Generated By Doctor Or AI

Beyond quantitative metrics, qualitative analysis and interpretability studies were conducted using tools such as LIME and SHAP to ensure the system's decisions are explainable and trustworthy—an essential requirement in the domain of healthcare applications. The explainability framework highlighted that the model tends to focus on specific markers such as empathetic phrasing, diagnostic specificity, and conditional advice, which are more prevalent in doctor-written responses. Conversely, responses flagged as AI-generated typically demonstrated overuse of generic phrasing, excessive formality, or templated sentence structures, which the model successfully identified as key distinguishing factors. This interpretability empowers patients and healthcare providers to understand and verify the rationale behind each classification, fostering greater trust in AI-assisted medical systems. Furthermore, the real-time response classification tool was integrated into a prototype web interface and tested in a simulated environment, wherein user interaction feedback confirmed the model's utility, usability, and transparency. Users appreciated the ability to receive not just a binary classification, but also a visual map of the linguistic cues that informed the model's decision. The results of this comprehensive evaluation demonstrate that MEDXNET is not only a high-performing model but also a viable candidate for real-world deployment in health consultation platforms. It represents a critical advancement in ensuring that AI is held accountable in sensitive domains like healthcare, where the accuracy and authenticity of information are paramount to patient safety and clinical decision-making.

VI. CONCLUSION

The increasing reliance on artificial intelligence in healthcare consultations presents both promising advancements and critical challenges. While AI can support scalability, efficiency, and access to information, its responses may lack contextual accuracy and human empathy, especially when trained on insufficient or biased datasets. To address this pressing issue, this study proposed MEDXNET, an efficient deep learning-based model designed to accurately classify whether a given medical consultation response was generated by a human doctor or an AI system. By integrating BiLSTM, transformers, and CNN1D layers with TF-IDF vectorization, MEDXNET captures both global semantic dependencies and localized textual nuances that distinguish AI-generated content from human-authored medical responses.Through rigorous testing on the custom-built MEDIC dataset, MEDXNET achieved a remarkable 95% accuracy, outperforming traditional models such as BiLSTM, GRU, and LSTM. An extended variant incorporating CNN2D layers further boosted performance to 96.78%, highlighting the effectiveness of multidimensional convolution in medical text classification. The use of interpretability tools like LIME and SHAP ensured transparency in decision-making, which is crucial for building trust in AI-assisted healthcare systems.This research makes a significant contribution toward responsible AI usage in medicine by empowering users with tools to verify the origin of their consultation responses. MEDXNET not only offers a high-performance classification framework but also sets the foundation for ethical AI deployment, where safety, accuracy, and explainability are paramount. As AI continues to evolve in the healthcare domain, systems like MEDXNET will be instrumental in bridging the gap between automation and accountability, ensuring that technological advancements align with patient trust and clinical integrity.

VII. REFERENCES

1. Vaswani et al. (2017) introduced the transformer model, which revolutionized natural language processing with its attention mechanisms.

2. Hochreiter and Schmidhuber (1997) proposed Long Short-Term Memory (LSTM), addressing vanishing gradient issues in RNNs.

3. Cho et al. (2014) developed the GRU model, an efficient alternative to LSTM for sequential data modeling.

4. Devlin et al. (2019) presented BERT, a transformer-based model pre-trained on large corpora, achieving state-of-the-art results in many NLP tasks.

5. Liu et al. (2019) optimized BERT into RoBERTa by training it longer and on more data, leading to further improvements in performance.

6. Rajpurkar et al. (2018) created the SQuAD 2.0 dataset to challenge models with unanswerable questions, enhancing model robustness.

7. Chen et al. (2021) provided a comprehensive review of deep learning applications in healthcare, highlighting both challenges and opportunities.

8. Ribeiro et al. (2016) introduced LIME, a method for explaining the predictions of any classifier, promoting transparency in AI systems.

9. Lundberg and Lee (2017) developed SHAP, a unified approach to interpreting model predictions based on cooperative game theory.

10. Esteva et al. (2019) discussed the potential of deep learning to transform healthcare delivery and diagnostics.

11. Yang et al. (2019) proposed XLNet, an autoregressive pre-training method that outperformed BERT on several benchmarks.

12. Jin et al. (2019) created PubMedQA, a dataset focused on biomedical question answering using expert-annotated content.

13. Zhang et al. (2020) improved biomedical embeddings through BioWordVec by integrating subword information and medical subject headings.

14. Johnson and Khoshgoftaar (2019) conducted a survey on handling class imbalance in deep learning, a key issue in healthcare datasets.

15. Turing (1950) questioned whether machines can think in his foundational paper on artificial intelligence and the Turing Test.