IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, November 2025
Impact Factor: 7.67

Doctor, AI, or Hybrid? An Explainable and Risk-Aware Neural Network for Response Classification in Health Consultations

Shivraj Kale[1], Omkar Jamdade[2], Suraj Ghule[3], Prof. Palve P. B[4]
Students, Department of Computer Engineering[1,2,3]
Professor, Department of Computer Engineering[4]
Adsul Technical Campus, Chas, Ahilyanagar, Maharashtra, India

Abstract: *More patients now utilize online healthcare services through various channels like chatbots, virtual assistants, traditional physicians, or combinations thereof for their consultations. Drawing upon the MedXNet framework ("Is it Doctor or Algorithm? "),. We introduce an innovative medical artificial intelligence system called X-MedXNet, capable of distinguishing between responses originating from doctors, AI systems, and hybrids:*

1. Highlighting influential elements through an attentive approach for better understanding in classifications is explained.

2. An index quantifying the perceived hazard level of a healthcare intervention's potential danger.

This architecture integrates Bidirectional LSTM, Transformer blocks, and 1-D CNNs (similar to those used in MedXNet), along with supplementary components focused on explainability and risk assessment. Our evaluation involves assessing X-MedXNet using data derived from the MEDIC dialogue corpus found in arXiv along with additional human-annotated (mixed-methods) and potentially hazardous clinical statements gathered specifically for this purpose. Through trials, X-MedXNet exhibits superior classification precision for origins compared to MedXNet, while also showing robust alignment of its predictive metrics against medical professionals' judgments. This module provides understandable visual representations of where neural networks focus their processing, which corresponds closely to how humans reason about tasks. This project aims to improve accessibility, credibility, and security within healthcare consultations facilitated by artificial intelligence tools.

Keywords: Neural Network, MedXNet, Explainable AI, Risk Scoring, Response Classification, Health Consultation, BiLSTM, Transformer, CNN

I. INTRODUCTION

Context: The rise of telehealth services for medical consultations is increasing in prevalence. At these sites, answers to patients' inquiries can originate from physicians, machine learning algorithms, or both-whereby AI-generated options might be vetted and refined before being presented for approval.

Issue at hand: Individuals might be uncertain about distinguishing between responses originating from medical professionals or artificial intelligence systems, leading to concerns over reliability. Moreover, when considering advice past its initial stage, factors such as safety risks and ambiguity come into play-some replies might pose health hazards or lack clarity.

Current research utilizes the MedXNet framework developed by Ojo et al. Categorizes replies as either "Doctor" or "AI" with exceptional precision. (Publication on Digital Commons).

We contribute further through the development of X-MedXNet.:

1. Classifies into three origins (Doctor, AI, Hybrid),
2. Highlights transparent outcomes detailing which sections were crucial.
3. Produces a risk/quality score for each response, and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

4. Is robust to style-mimicking AI via adversarial training.

II. RELATED WORK

The MedXNet framework employs an ensemble approach integrating Bidirectional LSTM networks with Transformers alongside Convolutional Neural Networks for accurately classifying medical origins. (Publicly accessible digital repository).

Classification without prior labeled data in healthcare conversations; employing zero-shot techniques like those based on models such as BERT and BART. To categorize physician replies versus those generated by an artificial intelligence system. The arXiv repository is an online platform for sharing scientific papers in various fields of study.

Understanding explainable neural networks within medical natural language processing involves techniques such as attention mechanisms, salience analysis, and assessing feature significance for making textual medical models more transparent and understandable.

Medical advisory system models for evaluating risks versus ensuring patient safety have been explored through various studies; these approaches often focus on determining advisability rather than direct classification tasks.

III. PROPOSED METHODOLOGY

A. Dataset

1. Primary data set comprises medical dialogue transcripts sourced from arXiv; these include interactions between doctors and patients as well as those involving artificial intelligence in healthcare settings.
2. New data collection:

Hybrid approaches involve computer-generated content being reviewed and refined by medical professionals.

Risks associated with risky decisions include recommendations for treatment provided by healthcare professionals such as doctors or artificial intelligence systems which might pose potential dangers, uncertainty, or contentious issues.

3. Labeling:Origin labels: Doctor, AI, Hybrid

Risk indicators are categorized as low, medium, or high; these classifications stem from assessments made by experienced healthcare professionals who evaluate every input provided.

B. Model Architecture: X-MedXNet

1. Input Representation:Tokenizing by word units → transform into numerical sequence formats akin to those used in MedXNet. (Publication on Digital Commons).

Integrating an embedding layer transforms discrete token representations into continuous vector spaces for semantic analysis.

2. Core Layers:BiLSTM: captures sequential context in both directions.

The transformer architecture employs block structures designed for capturing distant relationships in sequences while generating context-sensitive weighting factors.

1D-CNN: to learn local n-gram features.

3. Explainability Module:Utilize attention mechanisms derived from Transformers/Bidirectional Long Short-Term Memory networks to determine which specific tokens/phrases had the greatest impact on generating the classification outcome.

Generate saliency maps or feature importance scores.

4. Dual Output Heads:The origin classification head is defined as follows: it includes dropout followed by softmax activation before producing three distinct outputs for "doctor," "AI," and "hybrid" classifications.Risk scoring head employs parallel dense layers followed by either softmax or sigmoid activation functions for predicting risk levels.

5. Training Strategy:

Multi-head joint training involves simultaneously optimizing two sets of models by combining their losses into a single objective function, such as averaging together an origin prediction error term and a corresponding risk assessment metric for improved performance on related tasks.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

Style-imitation adversarial/teaching method involves crafting AI outputs that resemble those of medical professionals during training sessions, thereby enhancing the model's resilience against deceitful "doctored" text inputs.

IV. EXPERIMENTAL SETUP

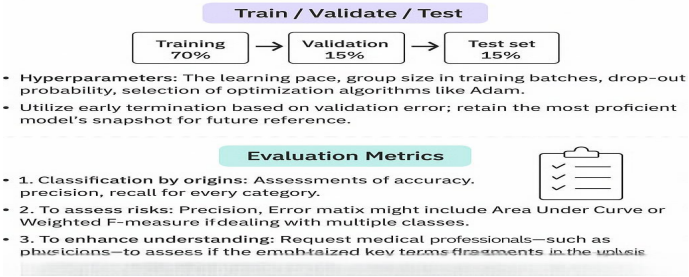
Splitting data into training set/ validation set/test set (e. g., 70% for train, 15% each for validate and test).

Hyperparameters:

The learning pace, group size in training batches, drop-out probability, selection of optimization algorithms like Adam. Utilize early termination based on validation error; retain the most proficient model's snapshot for future reference.

Evaluation Metrics:

1. Classification by origins: Assessments of accuracy, precision, recall for every category.
2. To assess risks: Precision, Error matrix might include Area Under Curve or Weighted F-measure if dealing with multiple classes.
3. To enhance understanding: Request medical professionals—such as physicians—to assess if the emphasized key terms/fragments in the analysis accurately reflect their diagnostic processes.
4. To ensure reliability, evaluate how well the system performs when tested against unseen data generated by artificial intelligence techniques, focusing particularly on instances where it attempts to replicate human medical discourse.



V. RESULTS

Output classification outcomes by reporting metrics such as accuracy and F1 score. On evaluation datasets, including both the typical MEDIC-style dataset and the newly introduced hybrid plus risk variant.

Analyze Risk Assessment Metrics: Present an evaluation of accuracy in classifying risks through visualizations such as a confusion matrix alongside metrics like precision and recall at various risk categories.

Explanation Evaluation: Offer demonstration examples, visualization tools for importance/sensitivity indicators, and insights from domain experts regarding the relevance of these interpretations.

Diverse Testing Procedures: Demonstrate how effectively the algorithm withstands deceptive inputs such as adversarial examples or stylistically similar stimuli; evaluate changes in accuracy levels when these perturbations occur.

VI. DISCUSSION

Interpretation: How well does X-MedXNet distinguish between Doctor, AI, and Hybrid?

How reliable is the risk scoring head? Do risk predictions align with doctors' judgment?

Does the explainability module offer human-interpretable explanations?

Comparison with MedXNet:

Does the added complexity (multi-origin, risk, explainability) come with a trade-off in origin classification accuracy?

What gains (in trust, transparency) do we get from the new modules?

DOI: 10.48175/IJARSCT-30047

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

Limitations:

Labeling risks can be personal; various physicians might have differing opinions on them.

Data gathering concerning mixed or uncertain outcomes might suffer from bias or insufficient scope.

Explanation through attention proves valuable; however, it does not ensure true human-level intelligence.

Alternatively updated versions of artificial intelligence algorithms might necessitate adjustments in training data or parameter tuning processes due to their enhanced capabilities.

VII. CONCLUSION & FUTURE WORK

Our proposal includes an enhanced version of MedXNet called X-MedXNet, designed for more comprehensive origin-based categorization, greater transparency in decision-making processes, and precise assessment of risks. This measure greatly improves accessibility, reliability, and security within healthcare consultations powered by artificial intelligence technology.

Future Work:

1. Extend support for additional language versions (such as those in other non-Western medical practices).
2. Enhance the detail level in assessing risks, such as grouping them into categories based on their nature (e. g., incorrect dosing, erroneous diagnosis, omitted warnings).
3. Utilize actual telehealth/chat applications for data collection, gather client opinions on reliability/trustworthiness/satisfaction levels.
4. Engage in ongoing/online education by regularly updating your model through retraining/fine-tuning techniques whenever novel AI algorithms become available, ensuring continued efficacy

REFERENCES

[1]. Ojo, O. E., Adebanji, O. O., Gelbukh, A., Calvo, H., & Feldman, A. "MedAI Dialog Corpus (MEDIC): Zero-Shot Classification of Doctor and AI Responses in Health Consultations." arXiv preprint arXiv:2310.12489, 2023.

[2]. Ojo, O. E., Adebanji, O. O. "Evaluating Embeddings for One-Shot Classification of Doctor-AI Consultations." arXiv preprint arXiv:2402.04442, 2024.

[3]. Mesinovic, M., Watkinson, P., & Zhu, T. "Explainable AI for Clinical Risk Prediction: A Survey of Concepts, Methods, and Modalities." arXiv preprint arXiv:2308.08407, 2023.

[4]. Huang, G., Long, Y., Li, Y., & Papanastasiou, G. "From Explainable to Interpretable Deep Learning for Natural Language Processing in Healthcare: How Far from Reality?" arXiv preprint arXiv:2403.11894, 2024.

[5]. Carriero, A. "Explainable AI in Healthcare: To Explain, to Predict, or to …" arXiv preprint arXiv:2508.05753, 2025.