2024 27th International Conference on Computer and Information Technology (ICCIT)
20-22 December 2024, Cox's Bazar, Bangladesh

Explainable AI-Driven Improved Disease Prediction through Symptom Analysis with Custom BERT

Md. Minhazul Islam*, Md. Tanbeer Jubaer†, Azmain Yakin Srizon‡, Md. Ali Hossain§,
Md. Farukuzzaman Faruk¶, S. M. Mahedy Hasan‖, A. F. M. Minhazur Rahman** and Nishat Tasnim Esha††
*†‡§¶‖** Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh
†† University of Rajshahi, Rajshahi, Bangladesh
E-mails: *minannu.2001@gmail.com, †tanbeerjubaer@gmail.com, ‡azmainsrizon@gmail.com, §ali.ruet@gmail.com,
¶faarukuzzaman@gmail.com, ‖mahedycseruet@gmail.com, **m.r.saurov@gmail.com, ††eshastat16@gmail.com

***Abstract*—This study proposes a custom BERT-based hybrid model aimed at enhancing the accuracy and explainability of disease prediction based on patient symptom descriptions. Utilizing the Symptom2Disease dataset, which includes 1,200 samples across 24 disease categories, the model processes textual symptom data and maps it to corresponding diseases. Key contributions of this work include the integration of Explainable AI (XAI) techniques to provide transparency in the model's decision-making process and the employment of dropout to mitigate overfitting, ensuring that the model generalizes well on unseen data. The proposed model builds upon BERT's pre-trained architecture by incorporating additional layers that refine the embeddings and enhance classification accuracy. Through rigorous experimentation, the model achieved an impressive 99% accuracy, outperforming existing models such as DistilBERT and MCN-BERT in both accuracy and training efficiency. The inclusion of XAI allows for visualizations of word-level contributions, offering insights into how specific symptoms influence disease predictions. Furthermore, a comparative analysis with previous studies demonstrates that the proposed model not only performs with higher accuracy but also achieves faster training times, making it both computationally efficient and effective for practical medical applications. The paper highlights the potential for applying hybrid architectures in healthcare, improving disease prediction through advanced natural language processing techniques.**

***Index Terms*—Disease Prediction, BERT, Explainable AI, Symptom2Disease, Medical Diagnosis, Text Classification, Model Interpretability, Overfitting Mitigation**

I. Introduction

Disease is a prevalent condition affecting all living organisms. Effective management of diseases and their treatments is essential, which is why individuals frequently seek medical attention. In medical science, each disease is associated with multiple symptoms. Some of these symptoms are unique and distinct, while others may be common across various diseases[1]. For a physician to provide appropriate treatment, it is crucial to accurately diagnose the patient's condition. To this end, the physician may order diagnostic tests to gain a clearer understanding. After analyzing the test results and patient-reported symptoms, the physician determines whether the patient is afflicted with a particular disease. In this diagnostic process, physicians rely on recognizing patterns in the presented symptoms and other relevant information to make informed decisions[1].

The objective of this experiment is to uncover these latent patterns from the descriptions provided by patients and predict potential diseases. This prediction system aims to assist physicians in making more accurate diagnoses. However, accurately predicting diseases is often a complex task, as natural language is inherently diverse and does not always follow clear, consistent patterns. In this study, efforts are made to effectively preprocess textual data and utilize advanced models capable of capturing intricate patterns in the data to predict diseases. Additionally, the study aims to provide insights into how and why the model arrives at specific disease predictions based on the text. By explaining the entire process, this research hopes to pave the way for future large-scale applications, ultimately enhancing the support provided to physicians in disease diagnosis.

II. Literature Review

High-quality data is crucial for achieving superior accuracy and performance in machine learning (ML) models. A study [2] emphasizes the significance of clean, precise, consistent, and complete datasets, demonstrating how data quality directly affects model performance and generalization. It highlights that improving data quality can reduce bias, enhance prediction accuracy, and boost the robustness of ML models. Another investigation [3] examines the relationship between six key data quality dimensions and the performance of fifteen popular ML algorithms across tasks such as classification, regression, and clustering. Through extensive experiments on real-world datasets, the researchers identified data completeness, feature accuracy, and target accuracy as the most critical factors influencing algorithm performance. A third paper [4] addresses the ethical implications of data quality in supervised machine learning, suggesting that behavioral data quality substantially impacts model behavior. It advocates for a shift from purely technical assessments of data quality to approaches that include ethical considerations, reinforcing the principle of "garbage in, garbage out" for data-intensive applications like supervised machine learning.

Recent advancements in hybrid deep learning models have contributed to substantial improvements in accuracy for various tasks. A study [5] presents an ensemble hybrid model that integrates RoBERTa with LSTM, BiLSTM, and GRU, effec-

979-8-3315-1909-4/24/$31.00 ©2024 IEEE

tively capturing long-range dependencies in text embeddings. This model outperforms state-of-the-art methods on datasets such as IMDb and Twitter US Airline Sentiment. Another research paper [6] proposes a robust text classification model combining RNN with BERT, achieving enhanced accuracy on the SST-2 dataset. In the context of Bangla sentiment analysis, a study [7] leverages BERT's transfer learning capabilities, proposing a CNN-BiLSTM model with BERT embeddings that significantly improves performance, achieving 94.15% accuracy on binary classification tasks. Other studies [8], [9] further support the efficacy of hybrid models, particularly those combining BERT-based models with BiLSTM and Bi-GRU layers, and the introduction of GBERT for fake news detection highlights the potential for hybrid architectures to outperform standalone models by leveraging deep contextual understanding and generative capabilities.

Various studies have explored deep learning approaches, such as BERT, for Named Entity Recognition (NER) in symptom extraction tasks. Research [10], [11] highlights BERT's effectiveness in understanding context, enabling accurate classification of symptoms from clinical notes and social media data. A study [12] leveraging transfer learning with BERT on a Bengali symptom-based dataset achieved 93.75% accuracy, contributing to medical diagnosis. Another study [13] applied NLP algorithms to classify Serious Illness Conversations (SIC) within EHR data using models like LR, XGBoost, and Bio+Clinical BERT, achieving high F1 scores for prognosis and goals classification. A comparative study [14] found that Random Forest outperformed other models, achieving 99.5% accuracy in disease prediction based on 132 symptoms. Further research [15] optimized classifiers like SVM, RF, and KNN for symptom-based disease prediction, with KNN achieving over 99% accuracy. Another study [16] examined MCN-BERT and BiLSTM models for disease prediction, with MCN-BERT achieving 99.58% accuracy, underscoring deep learning's potential in enhancing early diagnosis and remote healthcare.

Explainability in AI, particularly in healthcare, has gained significant attention. Research efforts have focused on enhancing model transparency and trust, especially in critical areas like healthcare and finance. Various studies [17], [18] emphasize the need for Explainable AI (XAI) in healthcare, suggesting that embedding explainability into models from the onset enhances trust in AI systems. Another study [19] reviews methods for providing natural language explanations for AI predictions, improving interpretability in sentiment analysis and fact-checking. Additionally, [20] categorizes explainability methods such as LIME and SHAP, emphasizing their importance in deep learning models to improve transparency and regulatory compliance.

The studies by [21], [16] both worked with the Symptom2disease dataset [22], achieving 93.33% and 99.58% accuracy, respectively. However, the latter study [16] employed expensive pre-training methods such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Neither study incorporated Explainable AI, which represents a significant research gap. This study aims to address these gaps by proposing an approach that incorporates both advanced deep learning techniques and Explainable AI for disease prediction.

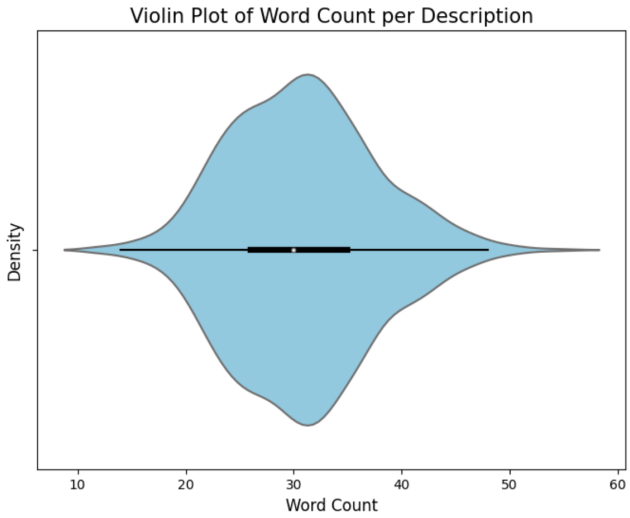

Fig. 1: Violin plot showing the distribution of text length across the dataset

III. Materials and Methods

A. Dataset Description

The dataset used in this study is sourced from Kaggle and is titled "Symptom2Disease" [22], comprising 1,200 samples across 24 common diseases. Each sample contains two columns: a textual description of symptoms and the corresponding disease label. The dataset is well-balanced, with 50 instances per disease class, making a total of 1,200 samples ($24 \times 50 = 1{,}200$). Although the dataset is clean and well-formatted, it poses some limitations. Given the large number of classes, the relatively small sample size makes it challenging to develop a robust model. Additionally, the descriptions are often brief, focusing on key features, which may not fully represent the complexity of real-world symptoms. In practical scenarios, patients often provide detailed accounts that may include irrelevant information, which can affect model performance. These challenges will be addressed during experimentation to achieve optimal results.

B. Data Preprocessing

As the dataset is pre-existing and has been widely used in prior research, there is limited scope for extensive data preprocessing. However, the disease labels are in text format, requiring conversion to numeric values for machine learning tasks. This is accomplished by mapping text labels to numeric values using a dictionary, which facilitates compatibility with most machine learning algorithms. Additionally, extra spaces are removed, and all text is converted to lowercase to enhance generalizability by reducing inconsistencies in case sensitivity. Unicode normalization is also applied to ensure consistent text representation, as varying fonts could impact results

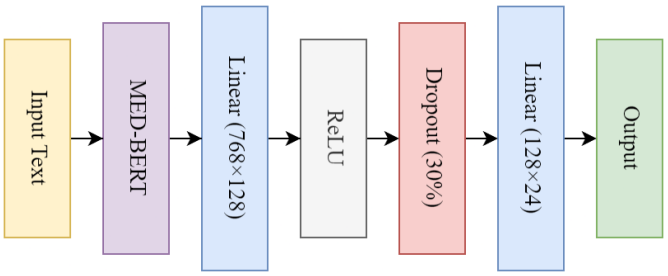

Fig. 2: Overview of the proposed BERT-based hybrid model architecture

and lead to misinterpretation of characters during processing. Furthermore, to ensure that text length variability does not affect the model's performance, a word count analysis was conducted. Figure 1 presents a violin plot illustrating the distribution of word counts across the dataset. The plot reveals that most descriptions contain 20 to 40 words, with the median word count approximately 30. The symmetrical shape of the plot indicates a balanced word count distribution with minimal outliers. The interquartile range (IQR) is relatively narrow, suggesting that the dataset consists of uniformly sized descriptions, contributing to overall consistency in textual content, which is beneficial for model training and evaluation.

C. Data Partitioning

After preprocessing, the dataset is divided for training and testing purposes. It is first shuffled and then split into training and test sets, with 80% allocated for training and 20% for testing. Tokenization is performed using the AutoTokenizer class, and the resulting dataset includes 'input_ids', 'token_type_ids', 'attention_mask', and 'label'. For training purposes, only 'input_ids' and 'attention_mask' are used as inputs for the model. The training and test sets are managed using a dataset dictionary to streamline the workflow.

D. Model Architecture

Given the small dataset with 24 distinct labels, a BERT-based hybrid model is proposed to effectively capture the complexity of the text. The model architecture is shown in Figure 2. In this architecture, the input text is passed through a BERT model, which outputs an embedding represented by a 768-dimensional vector. Specifically, the average of the last hidden state is taken as the embedding:

$$h_{avg} = \frac{1}{n}\sum_{i=1}^{n} h_i \tag{1}$$

where h_i represents the hidden state of the i-th token, and n is the total number of tokens. This 768-dimensional vector is then passed through a linear layer to reduce the dimensionality to a 128-size vector:

$$z = W_1 h_{avg} + b_1 \tag{2}$$

where W_1 and b_1 are the weight matrix and bias term of the linear layer. To introduce non-linearity and allow the model to capture complex patterns, a ReLU activation function is applied:

$$\text{ReLU}(z) = \max(0, z) \tag{3}$$

The ReLU output is further processed by a dropout layer, which randomly drops 30% of the parameters to mitigate overfitting. Initial experiments without dropout resulted in significant overfitting, where the model performed well on the training data but poorly on the test data. The inclusion of dropout helps prevent overfitting, making the model more robust. Finally, the output from the dropout layer is passed through another linear layer, producing a 24-dimensional output vector o, corresponding to the logits for the 24 disease classes:

$$o = W_2\text{ReLU}(z) + b_2 \tag{4}$$

The final disease prediction is determined by selecting the class with the highest logit value, which corresponds to the predicted disease class.

TABLE I: Experimental Settings for Model Training

Parameter	Value
Epochs	26
Learning Rate	2e-5
Train Batch Size	32
Evaluation Batch Size	8
Epsilon	1e-8
GPU	Kaggle T4X2 GPU

IV. Experimental Results

1) Model Initialization: For the classification task, a pre-trained BERT model trained on a vast medical text corpus was employed. Specifically, the "SapBERT-from-PubMedBERT-fulltext" model developed by the Cambridge Language Technology Lab at the University of Cambridge was utilized [23]. This model, referred to as Med-BERT, was integrated into the custom architecture. The model was then structured as previously outlined in the methodology. For training purposes, the Hugging Face `Trainer` class was used to streamline model optimization and evaluation.

A. Configuration Parameters

After selecting the model, the environment was configured for model training and evaluation.All hyperparameters are described in Table I. The training process was conducted for 26 epochs, using a learning rate of 2×10^{-5}. The training batch size was set to 32, while the evaluation batch size was smaller, with 8 samples per batch to efficiently handle model validation. An epsilon value of 1×10^{-8} was employed to prevent division by zero during the optimization process. The training was performed on a Kaggle T4X2 GPU, which ensured efficient handling of the computational load for both training and evaluation phases.

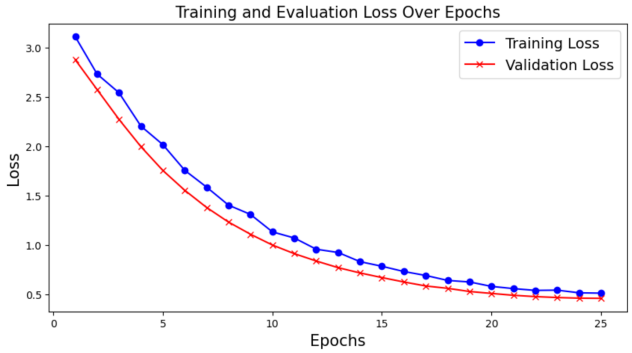

Fig. 3: Training and validation loss observed across epochs.

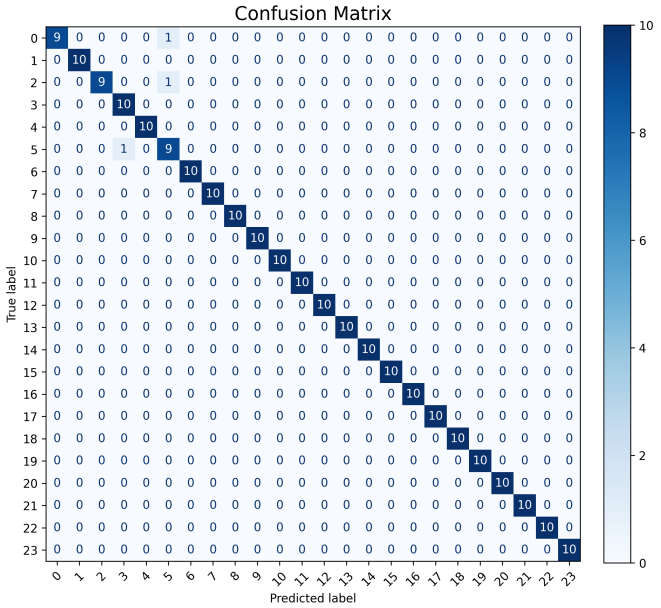

Fig. 4: Confusion matrix illustrating model performance across all classes.

TABLE II: Evaluation Metrics for Model Performance

Metrics	Value (%)
Average Precision	99
Average Recall	99
Average F1 Score	99
Accuracy	99

B. Obtained Results

Figure 3 illustrates the training and validation loss curves for our model over 26 epochs. As the number of epochs increases, the validation loss shows a consistent downward trend, indicating that the model's performance improves with more training. Following the completion of training, an accuracy of 99% was achieved. The average precision, recall, and F1 score were also high, reflecting the model's strong performance across all classes. The confusion matrix in Figure 4 provides detailed insight into the model's classification performance.

The results demonstrate that the model performs exceptionally well on this dataset. Multiple experiments were conducted with varying learning rates and epochs, and the settings described above yielded the best results. The model accurately classifies nearly all classes, with only a few misclassifications in isolated cases. Dropout was applied to mitigate overfitting, which significantly improved model generalization. In contrast to previous models where regularization techniques were not utilized, resulting in overfitting, the inclusion of dropout in this experiment effectively addressed the issue.

V. Result Analysis and Discussion

A. Model Performance Analysis

The model's performance depicted in Table II, reflected by a 99% accuracy, may seem excessively high, raising questions about its reliability. In this section, the results are analyzed, offering deeper insights and observations. Several factors, such as the dataset's structure, the model's architecture, and appropriate training, play a crucial role in these outcomes [24]. Since the dataset was sourced from Kaggle, there was no opportunity to modify its structure. Despite some preprocessing, no significant modifications were made to the data itself. Observations reveal that the dataset is well-organized and properly formatted, containing distinguishable features that facilitate the model's ability to capture patterns effectively. In real-world scenarios, datasets of this quality are rare, but rigorous data cleaning can often yield optimal results. Thus, the structured nature of the dataset likely contributes to the model's high accuracy.

Another factor influencing performance is the model architecture. Pre-trained BERT models generally outperform traditional state-of-the-art models in sequence classification tasks [25]. This approach was extended by proposing a custom hybrid model that includes additional layers beyond BERT, as discussed in the model description. Overfitting, a common issue with small datasets where the model excels in training but falters in testing, was mitigated through techniques such as dropout. Experimentation with various configurations helped address this overfitting problem, ultimately leading to the finalized model for classification [26].

A natural question arises: how does the model learn patterns from the dataset and predict outcomes? Machine learning models are often considered black boxes, where inputs are processed, and outputs are produced. Through the iterative optimization of a loss function, the model learns progressively, eventually reaching its optimal state. In the following sections, visualizations will illustrate how the model learns patterns and provides its predictions.

B. Training Data Insights

Visualization can be particularly useful in understanding how a model learns. For the training samples, the gradient of each word within the input sentences was calculated. Words that significantly influence the model's prediction for a given class exhibit higher gradient values. The LIME explainer was employed to compute and visualize these gradient values.

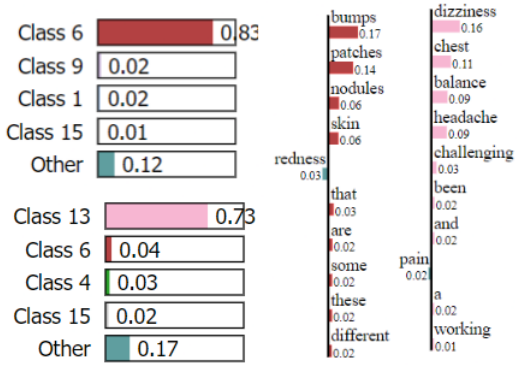

Fig. 5: Word probability distribution in training data.

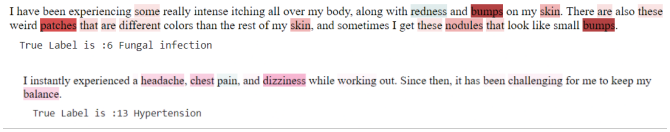

Fig. 6: Word contributions to final prediction during training.

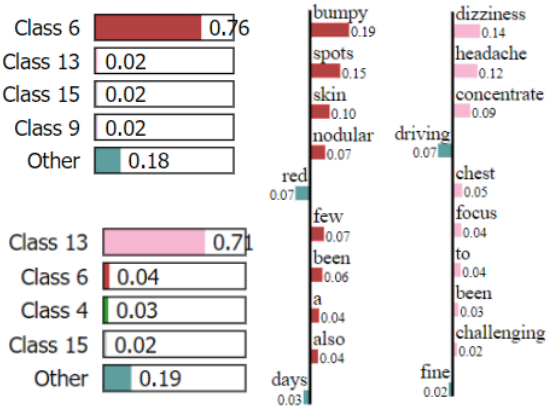

Fig. 7: Word probability distribution in test data.

The visualizations in figure 5, 6 indicate that the model successfully identifies the relevant words within sentences for accurate classification. For instance, in medical terminology, when detecting a fungal infection, the model highlights terms such as “bumps”, “patches”, and “nodules”, which are key indicators of the condition.

Another class, hypertension, was also evaluated, where the model emphasized terms like “dizziness”, “headache”, “chest”, alongside additional terms such as “balance” and “challenging”. These words align with common symptoms of hypertension, demonstrating that the model accurately identifies relevant terms for proper classification. This visualization confirms that the model focuses on meaningful features during the training process.

C. Test Data Insights

The test sample visualizations in figure 7, 8 demonstrate that the model consistently highlights relevant words, mirroring

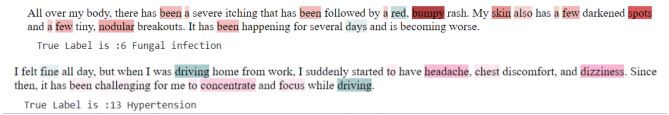

Fig. 8: Word contributions to final prediction during testing.

its behavior during training. For instance, in detecting fungal infections, the model focuses on terms such as “bumpy,” “spots,” and “skin nodular,” which are consistent with the relevant terms identified in the training data. Although the sentences differ, the model’s attention remains focused on pertinent medical descriptors, indicating a strong understanding of disease-related features.

Similarly, for hypertension, the model emphasizes words like “dizziness,” “headache,” “concentrate,” and “focus.” These words were similarly highlighted during training, affirming that the model generalizes its learning effectively across unseen test data.

The overall consistency in highlighting medically relevant terms across both training and test samples underscores the robustness of the model’s learning process. The accuracy of its predictions can be attributed to its ability to consistently capture key features.

D. Comparative Evaluation

This section presents a comparative analysis of studies that utilized the [22] dataset. The study conducted by [21] employed DistilBERT, which achieved an accuracy of 93.3%, but did not incorporate Explainable AI (XAI). In comparison, [16] implemented Medical Concept Normalization BERT (MCN-BERT), resulting in a higher accuracy of 99%. However, their approach relied on resource-intensive pre-training methods, such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), which significantly increased the overall training time. Specifically, the training of their classifier took 670 seconds, excluding the time required for MLM and NSP.

In contrast, the proposed model, a custom BERT with an additional classifier layer, achieved a comparable accuracy of 99% while reducing the training time to 333 seconds. A comparison of the confusion matrices shows that the proposed model maintains consistent performance across all classes. Moreover, by integrating Explainable AI, this model provides valuable insights into the decision-making process, offering a more comprehensive and transparent evaluation than previous studies.

Table III presents a comparison of recent studies, highlighting the variations in methodologies, results, and the integration of Explainable AI. This comparison underscores the advantages of the proposed model, particularly regarding accuracy, computational efficiency, and transparency.

E. Limitations

While the model demonstrates strong performance, certain limitations persist. The primary limitation is the relatively

TABLE III: Comparative Analysis of Studies Utilizing the Dataset [22]

Study	Methodology	Results	XAI
[16]	MCN-BERT + AdamP	99.58%	No
	MCN-BERT + AdamW	98.33%	
	Bidirectional LSTM + Hyperopt	97.08%	
[21]	DistilBERT	93.3%	No
	Ensemble Classification	91.7%	
Proposed Model	**Custom BERT**	**99%**	**Yes**

small size of the dataset, which can lead to overfitting, where the model becomes overly tailored to the training data and may struggle with generalization when presented with unseen data. Although the incorporation of dropout helped mitigate this issue, utilizing larger and more diverse datasets would enhance the model's robustness and improve its generalization capabilities.

VI. Conclusions and Future Directions

The conducted experiment demonstrated that a small, well-structured, and distinguishable dataset can lead to high performance, as reflected in the results. However, ensuring that the model performs optimally requires understanding both the dataset and the reasons behind the model's ability to recognize patterns. A model functions as a pattern recognizer, and thus, creating distinct features in both the dataset and the model is essential for improving performance. Additionally, close attention must be paid to the training process, including the visualization of key elements such as self-attention mechanisms and output gradients, to verify that the model is learning correctly. If the model fails to capture essential features, modifications in the dataset, model architecture, or loss function will be necessary to guide the model toward identifying the appropriate features. Future research will focus on exploring more complex and noisy datasets to evaluate the model's ability to make accurate predictions under diverse conditions. Further improvements will include incorporating additional features such as medical test summaries, diagnostic images, and other attributes into a multi-modal model to enhance prediction accuracy. Moreover, leveraging advanced techniques, including large language models (LLMs), will enable more effective handling of real-world medical diagnosis scenarios.

References

[1] P. Kumar and M. Clark, *Kumar Clark's Clinical Medicine*. Elsevier, 9th ed., 2016.

[2] S. Rangineni, "An analysis of data quality requirements for machine learning development pipelines frameworks," *International Journal of Computer Trends and Technology*, vol. 71, p. 16–27, Aug. 2023.

[3] L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, "The effects of data quality on machine learning performance," *arXiv preprint arXiv:2207.14529*, 2022.

[4] T. Hagendorff, "Linking human and machine behavior: A new approach to evaluate training data quality for beneficial machine learning," *Minds and Machines*, vol. 31, p. 563–593, Sept. 2021.

[5] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment analysis with ensemble hybrid deep learning model," *IEEE Access*, vol. 10, p. 103694–103704, 2022.

[6] C. Eang and S. Lee, "Improving the accuracy and effectiveness of text classification based on the integration of the bert model and a recurrent neural network (rnn_bert_based)," *Applied Sciences*, vol. 14, p. 8388, Sept. 2024.

[7] N. J. Prottasha, A. A. Sami, M. Kowsher, S. A. Murad, A. K. Bairagi, M. Masud, and M. Baz, "Transfer learning for sentiment analysis using bert based supervised fine-tuning," *Sensors*, vol. 22, p. 4157, May 2022.

[8] A. S. Talaat, "Sentiment analysis classification system using hybrid bert models," *Journal of Big Data*, vol. 10, June 2023.

[9] P. Dhiman, A. Kaur, D. Gupta, S. Juneja, A. Nauman, and G. Muhammad, "Gbert: A hybrid deep learning model based on gpt-bert for fake news detection," *Heliyon*, vol. 10, no. 16, 2024.

[10] X. Luo, P. Gandhi, S. Storey, and K. Huang, "A deep language model for symptom extraction from clinical text and its application to extract covid-19 symptoms from social media," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, p. 1737–1748, Apr. 2022.

[11] S. Azizi, D. B. Hier, and D. C. Wunsch II, "Enhanced neurologic concept recognition using a named entity recognition model based on transformers," *Frontiers in Digital Health*, vol. 4, Dec. 2022.

[12] M. M. Hossain, M. A. Mou, and M. N. N. Oishi, "Symptoms based disease prediction from bengali text using transformer network based pretrained model," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pp. 575–580, IEEE, 2022.

[13] A. Davoudi, H. Tissot, A. Doucette, P. E. Gabriel, R. Parikh, D. L. Mowery, and S. P. Miranda, "Using natural language processing to classify serious illness communication with oncology patients," *AMIA Summits on Translational Science Proceedings*, vol. 2022, p. 168, 2022.

[14] A. Das, D. Choudhury, and A. Sen, "A collaborative empirical analysis on machine learning based disease prediction in health care system," *International Journal of Information Technology*, vol. 16, no. 1, pp. 261–270, 2024.

[15] A. Fuster-Palà, F. Luna-Perejón, L. Miró-Amarante, and M. Domínguez-Morales, "Optimized machine learning classifiers for symptom-based disease screening," *Computers*, vol. 13, no. 9, p. 233, 2024.

[16] E. Hassan, T. Abd El-Hafeez, and M. Y. Shams, "Optimizing classification of diseases through language model analysis of symptoms," *Scientific Reports*, vol. 14, no. 1, p. 1507, 2024.

[17] S. Banerjee and D. Tomás, "Editorial: Explainable ai in natural language processing," *Frontiers in Artificial Intelligence*, vol. 7, Aug. 2024.

[18] T. Hulsen, "Explainable artificial intelligence (xai): Concepts and challenges in healthcare," *AI*, vol. 4, p. 652–666, Aug. 2023.

[19] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, and F. A. Batarseh, "Rationalization for explainable nlp: a survey," *Frontiers in Artificial Intelligence*, vol. 6, Sept. 2023.

[20] A. Diwali, K. Saeedi, K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "Sentiment analysis meets explainable artificial intelligence: A survey on explainable sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 837–846, 2024.

[21] R. Sarkar, A. Hossain, and A. Z. Ifti, "Language model-based deep learning for automated disease prediction from symptoms," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6, IEEE, 2023.

[22] N. R. Barman, F. Karim, and K. Sharma, "Symptom2disease." https://www.kaggle.com/datasets/niyarrbarman/symptom2disease, 2023.

[23] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Online), pp. 4228–4238, Association for Computational Linguistics, June 2021.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436–444, May 2015.

[25] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing bert against traditional machine learning models in text classification," *Journal of Computational and Cognitive Engineering*, vol. 2, p. 352–356, Apr. 2023.

[26] P. Charilaou and R. Battat, "Machine learning models and over-fitting considerations," *World Journal of Gastroenterology*, vol. 28, p. 605–607, Feb. 2022.