Language Model-based Deep Learning for Automated Disease Prediction from Symptoms

1st Ripa Sarkar
Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: ripa.sarkar@g.bracu.ac.bd

2nd Ahbab Hossain
Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: ahbab.hossain@g.bracu.ac.bd

3rd Akib Zabed Ifti
Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: akib.zabed.ifti@g.bracu.ac.bd

4th Shishir Kumar Das
Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: shishir.kumar.das@g.bracu.ac.bd

5th Md Humaion Kabir Mehedi
Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: humaion.kabir.mehedi@g.bracu.ac.bd

6th Abid Hossain
Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: abid.hossain@g.bracu.ac.bd

7th Ehsanur Rahman Rhythm
Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: ehsanur.rahman.rhythm@g.bracu.ac.bd

8th Annajiat Alim Rasel
Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
Email: annajiat@bracu.ac.bd

***Abstract*—Effective medical care depends on accurate and prompt disease prognosis based on patient symptoms. In this study, we suggest creating deep linguistics models that given a user's brief description of symptoms, can correctly predict the condition. We want to increase the effectiveness of disease prediction by utilizing deep learning technologies, leading to better patient outcomes. We demonstrate the efficacy and validity of our suggested techniques of disease prediction based on symptom reports through extensive trials and evaluation. A full grasp of the disease state and its corresponding symptoms is necessary in order to accurately predict disease from symptoms. Traditional methods frequently rely on manual diagnosis by medical experts which can be laborious and arbitrary. With the help of recent developments in deep learning, it is now possible to automate and enhance disease prediction based on symptom descriptions. In this paper, we suggest creating deep learning language models that, given a user's brief description of their symptoms can reliably predict illnesses. By learning from a lot of medical data, our model tries to identify complicated correlations between symptoms and diseases. By adopting deep learning architectures, we can leverage the ability of models to extract meaningful representations from text data and make accurate predictions. In our research paper, DistilBERT achieves the highest accuracy surpassing Ensemble classification with an accuracy rate of 93.3%.**

***Index Terms*—DL, DistilBERT, Ensemble, Random Forest, Gradient Boosting, Precision, Recall, F1-Score.**

I. Introduction

Clinical text data and other types such as surveys, reference papers are two categories for text data in smart healthcare. Human-robot treatment and patient-provider interactions depend on communication which is facilitated by tools like machine translation and user interfaces [1]. Effective clinician examination of patients' care choices is crucial to ensuring they receive treatment aligned with their values. It is vital to record patient-specific goals and prognostic data early in the trajectory of the illness. Assessment of shared decision-making, goal alignment and healthcare use is made easier by this documentation [3]. For decision-making, this documentation provide useful information [6]. But it is difficult to predict diseases from their symptoms and doing so requires a thorough understanding of medical problems and the symptomatology that goes along with them. Traditional methods frequently rely on manual diagnosis by medical staff which can be laborious and arbitrary. Using artificial intelligence to automate and improve disease prediction based on symptom descriptions is now possible because to recent advances in deep learning. In this paper, we suggest creating a deep learning language model that given a brief description of the user's symptoms can reliably identify diseases. By learning from a vast amount of medical data, our model tries to capture the intricate links between symptoms and diseases. We can take advantage of the model's capability by using a deep learning architecture.

II. Literature Review

Various NLP methodologies, their feature extraction methods and illustrative algorithms from past studies are compared in this study. The results show how well Neural NLP performs in Smart Healthcare. The document streamlines the choice of their own research algorithm by offering guidance on method selection for particular cases and a curated summary of pertinent research [1]. An NLP model is created by the researchers to extract data from CT scans on primary tumours and lymph nodes for multiturn question answering. ML techniques are used to develop prediction models for lymph node metastasis (LNM) using these extracted features and structured clinical data. The results demonstrate the superiority of RF models, achieving the best performance with an AUC value of 0.792 on the receiver operating characteristic curve [2]. The study introduces an NLP technique that automatically recognises and classifies SIC documents pertaining to prognosis and aims. Oncology patients' weakly labelled Electronic Health Record (EHR) data are used. The researchers are trained various ML algorithms to automatically categorise the data by domain and subdomain using a partially labelled dataset of SIC texts. SIC subdomains are identified and LR, XGBoost, BERT and Bio+Clinical BERT models are trained using comments from the training data [3]. The study deploys CUDoctor, a chatbot service within the Covenant University Doctor telehealth system, with the goal of assessing tropical disease symptoms in Nigeria. This chatbot employs user provided symptoms to predict diseases using fuzzy logic rules, fuzzy inference, NLP techniques and a fuzzy support vector machine. The system receives a strong usability rating of 80.4, affirming its effective personalized diagnostic capabilities [4]. In this work, NLP and ML are used to predict psychiatric symptoms and suicidal thoughts in recently discharged psychiatric patients from Madrid. Participants compare structured LR models to NLP models based on open-ended mood enquiries by providing consistent health assessments throughout time. In situations where surveys are difficult, NLP models demonstrate effectiveness in predicting symptoms and suicide ideation, providing quick, cost-effective identification of at-risk individuals. In situations where resources are scarce, NLP-driven techniques have the potential to be used as alternative data monitoring solutions [5]. Clinical narratives provide unique patient histories, assisting medical decisions. ML speeds up the development of NLP tools by using text data. In order to train clinical NLP models, this work meticulously evaluates text data's properties. By examining NLP goals and applications as well as data properties like size and source, it finds 110 pertinent studies using PubMed. Problems with data annotation are addressed, with recommendations for active learning and remote supervision for improved effectiveness [6]. While untapped clinical narratives frequently stay in EHRs, ML employing electronic health records (EHRs) reveals patient trajectories and disease risks. Text must be converted into structured data using NLP in order to allow for informed decisions and disease prevention. This study follows PRISMA requirements and investigates in-depth how NLP can be used to interpret clinical notes for long-term illnesses [7]. This study uses the need for glaucoma surgery using DL and data from the electronic health record (EHR). It includs both organised clinical data and clinical note. The patient demographics, diagnosis codes, surgeries and clinical measures are combined with the EHR ophthalmology notes. Initial notes are used to build ophthalmology-specific word embeddings. With the use of these embeddings and structured data, DL models correctly predicted surgery. AUC and F1 score are used in the evaluation; models employing free-text data outperformed models using solely structured data [8]. In this study, they introduce a method to identify diseases using a unique Bengali dataset focused on symptom-based predictions. By employing transfer learning, they build a disease prediction system using the BERT model, a transformer-based neural network. Through fine-tuning on their smaller dataset and utilizing advanced DL techniques, they achieve an impressive 93.75% accuracy in disease identification. This approach proves highly effective in using Bengali medical text and a transfer network-based pre-trained model to accurately predict relevant diseases based on symptoms [9].

III. Dataset Analysis

In Figure-1, the Symptom2Disease dataset is a collection of 1200 data points, each of which consists of a disease and its 50 associated symptoms [13]. The symptoms in the dataset are a mix of general and specific symptoms. Some examples of general symptoms include fever, fatigue and pain. Some examples of specific symptoms include cough, shortness of breath and rash. Different ML and DL models are used to predict diseases from symptoms. Such models are also used to identify potential diseases early on, allowing patients to seek medical attention and treatment promptly. Additionally, in situations where in-person consultations are not possible or desirable, the models can be used to provide remote diagnosis and treatment recommendations based on the user's symptoms. This dataset is a valuable resource for researchers and developers working on DL and ML models for disease prediction. The dataset is well-curated and easy to use, making it a good starting point for developing such models. Hence, there are some of the limitations of the Symptom2Disease dataset. The dataset is relatively small with only 1200 data points. This can limit the performance of ML models trained on the dataset. The symptoms in the dataset are not exhaustive. There may be other symptoms that are not included in the dataset. This can lead to false negatives where the model incorrectly predicts that a patient does not have a disease when they actually do. The dataset is not balanced. Some diseases are more common than others and this is reflected in the dataset. This can lead to the model being biased towards predicting the more common diseases. Despite these limitations, this is a valuable resource for researchers and developers working on ML models for disease prediction. The dataset is well-curated and easy to use, making it a good starting point for developing such models.

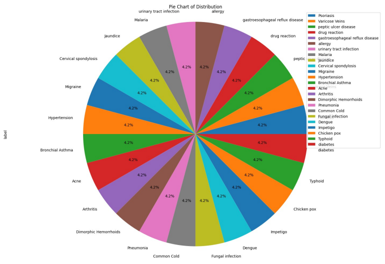

Fig. 1. Dataset Analysis Pie Graph.

In Figure-2 Analyzing the distribution of text lengths reveal variations in the number of words in the text on the dataset. The average text length is found to be approximately average length, with texts ranging from minimum length to maximum length words. Understanding this distribution is crucial for assessing the model's ability to handle texts of varying lengths.

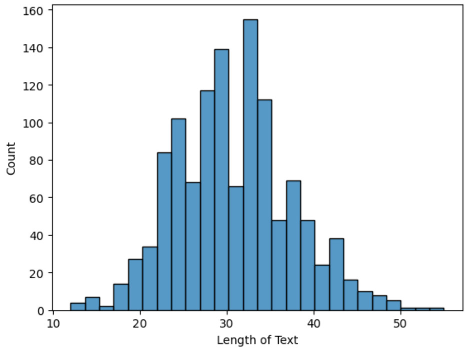

Fig. 2. Distribution of Text Lengths

IV. Methodology

We adopt a multi-step approach to develop our DL language model for disease prediction. It shows in Figure-3.

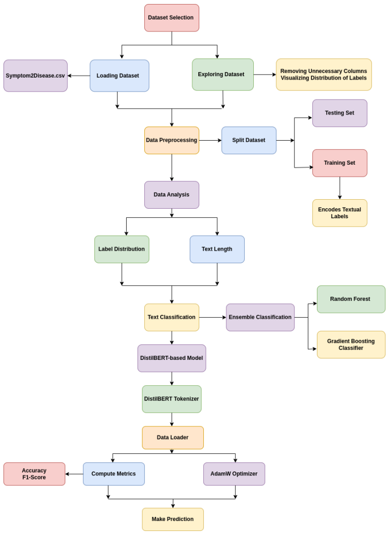

Fig. 3. Methodology

Data Collection: To initiate our study, we begin by loading the 'Symptom2Disease.csv' dataset which serves as the foundation for our research. This dataset is imported and stored in a DataFrame using Pandas, making it readily available for further analysis and modeling.

Data Preprocessing: As part of the data preprocessing steps, we address the need to encode textual labels into numerical values. This is accomplished using the LabelEncoder from the scikit-learn library, facilitating the model training process.

Feature Engineering: Our next crucial step involves preparing the text data for modeling. We harness the power of TF-IDF vectorization to convert the raw text into numerical features. The TF-IDF vectorizer is configured with a maximum of 1000 features although this parameter can be adjusted as needed. The resulting TF-IDF matrix (X) and labels (y) were ready for the subsequent modeling phase.

Data Splitting: To evaluate our models effectively, we divide the dataset into training and testing sets using the train_test_split function. This enables us to assess model performance on unseen data. The split ratio chosen is 80% for training and 20% for testing and we set a random seed (random_state=42) for reproducibility.

Model Initialization and Training: We embark on model training by initializing two base models: a Random Forest Classifier (rf_model) and a Gradient Boosting Classifier (gb_model). Both models are configured with a random seed (random_state=42) to ensure consistent results across runs. These models are then trained on the training data to learn the underlying patterns in the TF-IDF transformed text features.

Model Prediction: After training the base models, we make predictions on the test dataset. The predictions from the Random Forest (rf_predictions) and Gradient Boosting (gb_predictions) models are collected.

Ensemble Learning: In pursuit of enhancing prediction accuracy, we employ an ensemble learning approach. By combining the predictions of our base models through majority voting, we create an ensemble prediction. This is accomplished by stacking the individual predictions from the two base models and determining the most frequent prediction for each sample. The final ensemble predictions (final_predictions) thus represent the consensus of our base models.

Model Evaluation: To quantify the performance of our ensemble model, we calculate the accuracy, which measures the proportion of correct predictions in the test set. The ensemble achieved an accuracy of approximately 0.917, showcasing its effectiveness in classifying diseases based on symptoms.

V. Result Analysis

In this section, we present a comprehensive analysis of the results obtained from our text classification model trained to predict diseases based on symptoms. The analysis encompasses various aspects, including dataset distribution, model performance, and practical implications. Here we have implemented two types of model,

Firstly, we have trained a DistilBERT-based model for sequence classification to predict diseases from symptom descriptions. The model's performance is evaluated using various metrics, including accuracy and the F1 score. Here, we provide a detailed analysis of the model's performance in Figure-4:

epo.png

Fig. 4. Result Analysis of Model Performance

In Figure-4, the model is trained for 10 epochs using an AdamW optimizer with a learning rate of 5e-5. Training is conducted on a GPU platform.

The accuracy metric measures the proportion of correctly predicted labels. The model achieves an accuracy of final accuracy on the validation set indicating its ability to correctly classify diseases based on symptoms. The F1-score which balances precision and recall, reaches 100 on the validation set. This metric provides insight into the model's overall classification performance. Throughout the training process, the model demonstrates a steady improvement in both accuracy and the F1-score. However, it is noteworthy that the training accuracy reached 100% in the final epoch, suggesting potential overfitting. Careful consideration of validation metrics is necessary to ensure model generalization. So we can claim that our text classification model demonstrates promising results in disease prediction based on symptoms. It represents a valuable tool for assisting healthcare professionals in diagnosing diseases promptly. However, further research and validation are necessary to ensure its robustness and applicability in clinical practice. Here, we also provide the model's performance graph in Figure-5.

gra.png

Fig. 5. Model Performance Graph

In our evaluation, we also compute precision and recall for the DistilBERT-based model and it achieves perfect scores of 100% for both precision and recall. Precision and recall are fundamental metrics used in classification to assess a model's performance, particularly when dealing with imbalanced class distributions. Precision quantifies the accuracy of positive predictions, revealing how many predicted positives are accurate, making it crucial in scenarios where false positives are costly or undesirable such as medical diagnoses. Recall gauges the model's ability to identify all actual positive instances, indicating how many true positives are correctly identified and is vital in situations where missing positive instances carries significant consequences like in search and rescue operations. By considering both precision and recall, we gain a comprehensive view of the model's performance, highlighting its excellence in making accurate positive predictions and capturing all actual positive instances.

Secondly, we have performed an ensemble classification using two base models, a Random Forest Classifier (rf_model) and a Gradient Boosting Classifier (gb_model) to predict the labels of a dataset containing symptoms and diseases. After encoding the labels as integers and converting the text data to TF-IDF features, the models are trained and used to make predictions on a test dataset. The ensemble predictions are generated by combining the predictions of the two base models using majority voting. The resulting ensemble model achieves an accuracy of approximately 91.67% and an F1-score of approximately 0.917 on the test data. This indicates that the ensemble of RF and Gradient Boosting models performs well in classifying symptoms to diseases, providing a robust and accurate solution for the given classification task. However, it's essential to further evaluate the model's performance, consider potential overfitting and fine-tune hyper-parameters for optimal results in a real-world scenario. We also compute the precision and recall are 93.1% and 91.7% respectively.

TABLE I
PERFORMANCE TABLE.

Model Name	Accuracy	Precision	Recall	F1-Score
DistilBERT	0.933	1.00	1.00	1.00
Ensemble Classification	0.917	0.931	0.917	0.917

Two different approaches for classification tasks are demonstrated in Table-I. The first part of the code focuses on text classification using the DistilBERT model which is a state-of-the-art transformer model. After preprocessing the text data and encoding labels, the DistilBERT model is trained for multiple epochs. The training progress is displayed, showing the accuracy, precision, recall, F1-score, loss and other metrics on both the training and validation datasets during each epoch. The model appears to perform exceptionally well with a final training accuracy of 100%, precision of 100%, recall of 100% and a validation accuracy of approximately 93.3% after ten epochs. The second part of the code employs an ensemble of RF and Gradient Boosting classifiers on a dataset where symptoms are classified into diseases. This ensemble approach achieves an accuracy of approximately 91.67%, precision of 93.1%, recall of 91.7% and an F1-score of approximately 91.70%.

In comparing the two approaches, it's important to note that they address different types of classification tasks. The ensemble approach deals with structured data where symptoms are directly linked to diseases, while the DistilBERT approach tackles text-based classification, likely with a different dataset and task. Therefore, a direct comparison between the two is challenging, as they serve distinct purposes. The effectiveness of each approach should be evaluated in the context of their respective tasks and datasets, taking into account factors like data quality, model complexity, and computational resources required.

The decision to employ DistilBERT and Ensemble Classification techniques on our dataset which deals with symptoms and diseases, is driven by several key factors. Firstly, DistilBERT is chosen for its proficiency in handling complex textual data, capturing nuanced language, and understanding context, which aligns well with the nature of medical text analysis. Moreover, its computational efficiency makes it feasible for resource-constrained environments. Ensemble Classification combining Random Forest and Gradient Boosting is utilized to enhance model robustness and classification performance. This ensemble approach mitigates overfitting, aids generalization and proves valuable when working with potentially noisy or heterogeneous medical datasets. Overall, these techniques offer versatility, strong benchmarking potential and the prospect of achieving accurate and reliable predictions for a range of symptoms and disease-related tasks, making them well-suited for our dataset.

VI. Future Work

In the context of future development, this project presents several avenues for enhancement. One area to explore involves delving into alternative transformer architectures such as BERT, RoBERTa or ELECTRA to assess their impact on classification accuracy and model performance. Another crucial step is conducting a rigorous hyperparameter tuning process involving methods like grid search or Bayesian optimization to identify the optimal learning rate, batch size and other critical parameters that can significantly influence training outcomes. To counteract the risk of overfitting, it would be beneficial to implement advanced regularization techniques like dropout, L2 regularization or batch normalization ensuring that the model generalizes effectively to unseen data. Additionally, data augmentation can be applied to artificially diversify the training dataset, thereby enhancing the model's ability to handle variations in input text. Ensemble methods which combine the predictions of multiple models can be explored as a means of boosting performance by leveraging complementary strengths. Thorough error analysis is paramount in understanding model limitations and refining its predictive capabilities. By scrutinizing misclassified instances it's possible to gain insights into specific patterns or contexts that challenge the model's accuracy. Furthermore, an extension to multi-label classification could be considered if instances might pertain to multiple disease categories, allowing the model to capture such complex relationships. Taking a step into transfer learning, pre-training the model on a broader medical text corpus could serve as a valuable foundation followed by fine-tuning on the specific dataset. Attention mechanisms within the model architecture can also be investigated to interpret which portions of the input text are driving the classification decisions, yielding valuable insights for medical professionals and users alike. To address class imbalance, techniques like oversampling, undersampling or generating synthetic data can be applied to ensure that the model remains unbiased towards prevalent classes. Collaboration with medical experts is crucial for validation and aligning the model's predictions with established medical standards. From a user interaction perspective, developing an intuitive and user-friendly web interface where users can input symptoms and receive disease predictions could facilitate widespread adoption. Ensuring the security of user data during deployment is of paramount importance. Ultimately, the long-term vision involves potentially transitioning the model into clinical practice, necessitating alignment with regulatory standards and ethical guidelines to ensure its safe and responsible use within healthcare settings. This holistic approach to future work can empower the project to realize its full potential in aiding disease diagnosis and classification based on symptoms.

VII. Conclusion

In conclusion, this research endeavors to advance disease prediction from patient-reported symptoms using deep learning and natural language processing. Our aim is to enhance accuracy and efficiency in disease prognosis. Conventional methods rely heavily on manual diagnosis which is time-consuming and subjective. Leveraging recent strides in deep learning, we've developed a proficient language model capable of accurately identifying diseases from symptom descriptions. While the model shows promise with commendable precision, recall and F1 score of 100 on the validation set, overfitting remains a concern. Future work involves exploring alternative architectures, . The ultimate goal is the responsible integration of this technology into clinical practice with a commitment to ethics, data security and regulatory standards, offering the potential to improve patient care and diagnostic accuracy significantly.

References

[1] B. Zhou, G. Yang, Z. Shi & S. Ma. (2022). Natural Language Processing for Smart Healthcare. IEEE Reviews in Biomedical Engineering, DOI: 10.1109/RBME.2022.3210270.

[2] Hu, D., Li, S., Zhang, H., Wu, N., & Lu, X. (2022). Using Natural Language Processing and Machine Learning to Preoperatively Predict Lymph Node Metastasis for Non–Small Cell Lung Cancer With Electronic Medical Records: Development and Validation Study. JMIR Medical Informatics, 10(4), e35475. https://doi.org/10.2196/35475.

[3] A. Davoudi, H. Tissot, A. Doucette, P. E. Gabriel, R. Parikh, D. L. Mowery & S. P. Miranda. (2022). Using Natural Language Processing to Classify Serious Illness Communication with Oncology Patients. AMIA Jt Summits Transl Sci Proc. 2022; 2022: 168–177, PMCID: PMC9285137, PMID: 35854756.

[4] Omoregbe, N., Ndaman, I. O., Misra, S., Abayomi-Alli, O., & Damaševičius, R. (2020). Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic. Journal of Healthcare Engineering, 2020, 1–14. https://doi.org/10.1155/2020/8839524.

[5] B. L. Cook, A. M. Progovac, P. Chen, B. Mullin, S. Hou & E. B. Garcia. (2016). Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid. Computational and Mathematical Methods in Medicine, Volume 2016, https://doi.org/10.1155/2016/8708434.

[6] I. Spasic & G. Nenadic. (2020). Clinical Text Data in Machine Learning: Systematic Review. JMIR Medical Informatics, Vol 8 , No 3, doi:10.2196/17984.

[7] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi & V. Osmani. (2019). Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Medical Informatics, Vol 7 , No 2, doi:10.2196/12239.

[8] S. Y. Wang, B. Tseng & T. Hernandez-Boussard. (2022). Deep Learning Approaches for Predicting Glaucoma Progression Using Electronic Health Records and Natural Language Processing. Ophthalmology Science, Volume 2, Issue 2, https://doi.org/10.1016/j.xops.2022.100127.

[9] M. M. Hossain, M. A. Mou & M. N. N. Oishi. (2022). Symptoms Based Disease Prediction from Bengali Text Using Transformer Network Based Pretrained Model. International Conference on Computer and Information Technology (ICCIT), INSPEC Accession Number: 22724680, DOI: 10.1109/ICCIT57492.2022.10055374.

[10] N. A. I. Omoregbe, I. O. Ndaman, S. Misra, O. O. Abayomi-Alli & R. Damaševičius. (2020). Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic. Journal of Healthcare Engineering, Volume 2020, Article ID 8839524, https://doi.org/10.1155/2020/8839524.

[11] P. Ding, Y. Pan, Q. Wang & R. Xu. (2022). Prediction and evaluation of combination pharmacotherapy using natural language processing, machine learning and patient electronic health records. Journal of Biomedical Informatics, Volume 133, September 2022, 104164, https://doi.org/10.1016/j.jbi.2022.104164.

[12] N. Rezaii, P. Wolff & B. H. Price. (2022). Natural language processing in psychiatry: the promises and perils of a transformative approach. The British Journal of Psychiatry , Volume 220 , Issue 5 , May 2022 , pp. 251 - 253, DOI: https://doi.org/10.1192/bjp.2021.188.

[13] N. R. Barman, F. Karim & K. Sharma. (2023). Symptom2Disease. Available on https://www.kaggle.com/datasets/niyarrbarman/symptom2disease.