DIAGNOSURE: AN AI-POWERED HEALTHCARE SUPPORT SYSTEM

Nandita P Nair[1], Rohith T C[2], Sradha S[3], Sreeram C[4], Silja Varghese[5]

[1]*Bachelor of Technology in CSE, NCERC*
[2]*Bachelor of Technology in CSE, NCERC*
[3]*Bachelor of Technology in CSE, NCERC*
[4]*Bachelor of Technology in CSE, NCERC*
[5]*Assistant Professor, Department of CSE, NCERC*

Abstract - DiagnoSure is an Artificial Intelligence (AI)-powered healthcare support system designed to assist in preliminary disease prediction using patient-provided symptoms. The system addresses the challenge of processing unstructured medical inputs expressed in natural language. A domain-specific Natural Language Processing (NLP) model, BioBERT (Bidirectional Encoder Representations from Transformers for biomedical text), is utilized to extract relevant medical features from free-text symptom descriptions [1]. These features are processed using Extreme Gradient Boosting (XGBoost) to generate disease predictions along with associated probability scores [2]. To ensure transparency and interpretability, SHapley Additive exPlanations (SHAP) is integrated to identify the contribution of individual symptoms to prediction outcomes [7]. The system also supports structured inputs, enabling efficient processing in Electronic Health Record (EHR) scenarios [9]. The proposed approach demonstrates reliable and interpretable predictions and serves as a decision-support tool for early health assessment.

Key Words: Artificial Intelligence, Natural Language Processing, BioBERT, XGBoost, Explainable AI, Disease Prediction.

1. INTRODUCTION

The integration of Artificial Intelligence (AI) into healthcare has significantly improved the efficiency and accuracy of disease prediction and clinical decision support systems. Advanced machine learning and deep learning models are increasingly used to analyze medical data and assist in early diagnosis [7]. However, a major challenge lies in processing patient-provided symptoms, which are often expressed in unstructured natural language formats.

Traditional healthcare systems primarily rely on structured data formats and fail to capture the semantic and contextual meaning of free-text inputs [8]. Domain-specific language models such as BioBERT have been developed to address this limitation by learning contextual representations from biomedical corpora [1]. Similarly, ClinicalBERT enhances understanding of clinical narratives and improves disease classification tasks [11]. In addition to accuracy, interpretability has become a critical requirement in healthcare applications. Black-box models reduce trust and limit their adoption in clinical practice. Explainable AI techniques such as SHAP provide transparency by identifying feature contributions to predictions [2][7].

To address these challenges, DiagnoSure is proposed as an AI-powered healthcare support system that integrates BioBERT for text processing, XGBoost for prediction, and SHAP for interpretability. The system is further enhanced with real-world features such as hospital recommendations, appointment booking, urgency detection, and LLM-based explanations, making it a comprehensive healthcare solution.

2. PROBLEM STATEMENT

Existing healthcare systems depend heavily on structured data and lack the ability to effectively process unstructured symptom descriptions. They often provide generic outputs without interpretability and lack integration with real-world healthcare services. These limitations reduce system reliability and hinder early diagnosis.

3. LITERATURE REVIEW

Recent advancements in healthcare Artificial Intelligence highlight the importance of domain-specific NLP models and explainable machine learning techniques. BioBERT has demonstrated strong performance in biomedical text mining tasks [1], while ClinicalBERT improves disease classification using clinical narratives [11][13]. XGBoost is widely used for medical prediction due to its high accuracy and efficiency [2][10]. SHAP enhances interpretability by explaining model predictions [6][7]. Models such as MedBERT support structured electronic health record analysis [9]. However, existing approaches lack integration of prediction, explainability, and healthcare services within a single system.

Furthermore, many systems are limited to either structured or unstructured data processing, reducing their flexibility in real-world scenarios. The absence of user-centric features such as medical guidance, accessibility to

healthcare providers, and real-time assistance further limits their practical usability. These gaps highlight the need for a comprehensive system that combines accurate prediction, interpretability, and healthcare service integration.

Table -3.1: Literature review analysis

S. No.	Paper Name	Core Content Focus	Project Contribution
1.	BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining [1]	Domain-specific language model (BioBERT) for biomedical text	Feature Extraction: Bio-BERT used for accurate extraction of entities from clinical text.
2.	XGBoost, a Machine Learning Method, Predicts Spinal Neurological Recovery in Patients with Cervical Cord Injury [10]	Application of the XGBoost machine learning algorithm for clinical prediction.	Prediction Model: Adopt XGBoost as the high-performance core algorithm for the final diagnosis / prognosis.
3.	The Role of Explainable Artificial Intelligence in Disease Prediction: A Systematic Literature Review and Future Research Directions [7]	Systematic review on the necessity and methods of Explainable AI (XAI) in healthcare.	Model Interpretation: Implement XAI (e.g., SHAP, LIME) to explain why the model made a specific diagnosis.
4.	Accurate Medical Named Entity Recognition Through Specialized NLP Models [5]	High-accuracy method for Medical Named Entity Recognition (NER) using specialized NLP (like BioBERT).	Data Preprocessing: Justifies and guides the use of specialized models for highly accurate extraction of features from medical text.

4. METHODOLOGY AND SYSTEM FEATURES

The proposed system, DiagnoSure, is designed as an intelligent healthcare support platform that integrates Natural Language Processing (NLP), machine learning, and Explainable Artificial Intelligence (XAI) to provide accurate, interpretable, and user-centric medical assistance. The system follows a multi-stage pipeline consisting of feature extraction, disease prediction, explanation generation, and healthcare service integration.

4.1 System Overview

DiagnoSure is a web-based healthcare support system that integrates Natural Language Processing (NLP), machine learning, and explainable AI techniques. The system processes user inputs and generates interpretable predictions through a multi-stage pipeline.

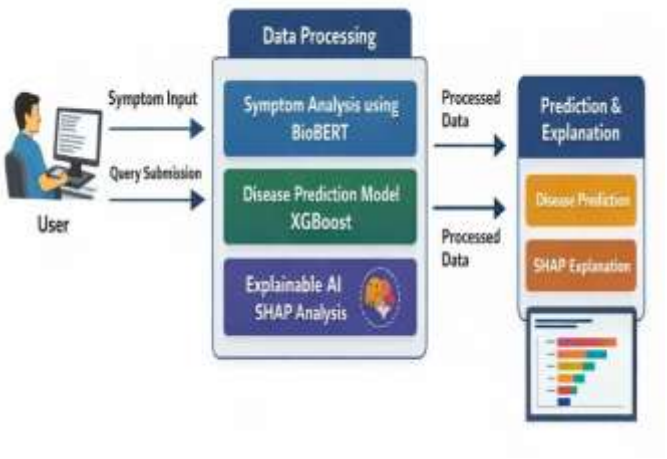

Fig - 4.1.1: System Architecture

The overall system architecture is illustrated in Fig. 4.1.1. The system consists of input processing, feature extraction using BioBERT, disease prediction using XGBoost, and explanation generation using SHapley Additive exPlanations (SHAP). The system also integrates additional modules such as report generation, recommendation system, and user interaction features.

4.2 Core Prediction Pipeline

The system processes inputs through a three-stage architecture:

- **Feature Extraction:** BioBERT is used to extract medical entities and contextual information from unstructured symptom descriptions. Domain-specific pretrained models such as BioBERT and ClinicalBERT have demonstrated superior performance in biomedical text understanding [1][11].
- **Disease Prediction:** Extracted features are passed to the Extreme Gradient Boosting (XGBoost) model, which generates disease predictions along with probability scores. These probability scores indicate the likelihood of each disease, improving decision-making. XGBoost is widely used in clinical prediction due to its robustness and ability to handle complex feature interactions [2][10].

- **Explainability:** SHapley Additive exPlanations (SHAP) is integrated to provide interpretability by identifying the contribution of each symptom to the prediction. Explainable AI techniques enhance transparency and trust in healthcare systems [6][7].

4.3 Dual Input Processing Mechanism

The proposed system supports a dual input processing mechanism to handle both unstructured and structured data efficiently. This design enables flexibility in processing different types of user inputs while maintaining accuracy and performance.

In the case of unstructured input, users enter symptoms in natural language. The system utilizes BioBERT to extract key medical features and contextual information from the input text. This process generates a context-aware representation of the symptoms, which is then passed to the XGBoost model for disease prediction. The model produces predictions along with probability scores indicating the likelihood of each condition. Subsequently, SHapley Additive exPlanations (SHAP) is applied to interpret the results by identifying the contribution of each symptom to the prediction. This approach leverages advanced Natural Language Processing (NLP) techniques to effectively handle free-text inputs and capture semantic meaning [5][8].

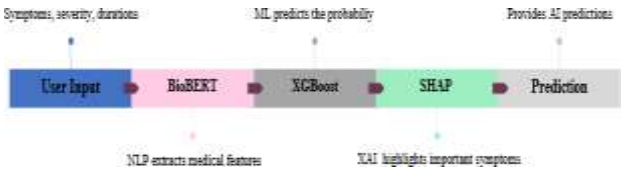

Fig. 4.3.1: Unstructured Input flow

For structured input, predefined medical data is directly provided to the system. In this case, the NLP stage is bypassed, and the input features are directly processed by the XGBoost model. The model generates predictions efficiently, and SHAP is used to provide feature importance and interpretability. This method ensures faster processing while maintaining accuracy. Such an approach aligns with structured Electronic Health Record (EHR)-based prediction systems, including models like MedBERT [9].

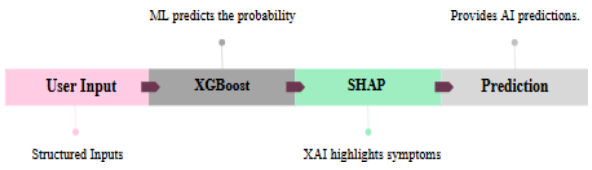

Fig. 4.3.2: Structured Input flow

4.4 Intelligent System Features

In addition to prediction, DiagnoSure incorporates several user-centric features that enhance usability, accessibility, and real-world applicability. The system maintains a user prediction history that stores previous inputs and corresponding results, allowing users to revisit past analyses and improving overall user experience. It also generates downloadable medical reports that include prediction results, XGBoost probability scores, and SHAP-based explanations, which can be used for future reference or clinical consultation.

The system further provides recommended precautionary measures based on the predicted condition, guiding users on initial steps before seeking medical attention. To support healthcare accessibility, it includes hospital and doctor recommendation functionality, which suggests appropriate healthcare providers based on the prediction. Additionally, an appointment booking system enables users to schedule consultations directly, reducing delays in accessing medical care.

Map integration is incorporated to display nearby hospitals and provide navigation support, improving accessibility to healthcare facilities. The system also includes an urgency detection mechanism that analyzes symptom severity and categorizes cases into different urgency levels, helping users prioritize medical attention. Furthermore, an LLM-based explanation system generates detailed natural language explanations of predictions, enhancing user understanding, improving transparency, and increasing trust in the system.

4.5 System Advantages

The proposed system offers several advantages by integrating advanced technologies into a unified framework. It combines Natural Language Processing and machine learning to handle both structured and unstructured data effectively. The system provides not only accurate predictions but also interpretability through explainable AI techniques. The inclusion of real-world healthcare services such as recommendations, appointment booking, and navigation significantly enhances usability. Overall, the system improves accessibility, supports early diagnosis, and enables informed decision-making.

4.6 System Requirements

The system requires both hardware and software components for effective implementation. From a hardware perspective, a system with an Intel i5 processor or higher, a minimum of 8 GB RAM, and at least 256 GB of storage is recommended to ensure smooth performance. Reliable internet connectivity is also necessary for real-time processing and interaction.

From a software perspective, the system can be implemented on operating systems such as Windows or Linux. Python is used as the primary programming language due to its strong support for machine learning and NLP libraries. The web application is developed using a lightweight framework such as Flask. A database system such as MySQL is used for data storage and management. Libraries including Transformers, and SHAP are utilized for model development and explainability.

4.7 Technologies Used

The system utilizes a combination of advanced technologies to achieve accurate and interpretable predictions. BioBERT is used for biomedical text processing and feature extraction [1], while ClinicalBERT enhances clinical text understanding [11]. XGBoost is employed as the primary prediction model due to its efficiency and ability to generate probability scores [2]. SHAP is used to provide explainable AI capabilities by identifying feature contributions [7].

In addition, Natural Language Processing techniques are applied to process unstructured inputs [8]. The frontend of the system is developed using web technologies such as HTML, CSS, and JavaScript, while the backend is implemented using the Flask framework for seamless integration of AI models and user interaction.

5. RESULTS AND DISCUSSIONS

The proposed system was evaluated using multiple user inputs to analyze its performance, usability, and effectiveness. The system successfully processes both structured and unstructured inputs and generates accurate and interpretable predictions. The system was evaluated using multiple real-time user inputs to assess its performance and usability.

5.1 User Login Interface

The system begins with a secure login interface that allows users to access personalized healthcare services.

Fig. 5.1: User Login Interface

The login mechanism ensures data privacy and secure interaction with the system.

5.2 User Input Interface

The system provides a simple and user-friendly interface that allows users to enter symptoms in natural language. The interface is designed to ensure ease of use and accessibility.

Fig. 5.2: User Input Interface

The input interface supports flexible symptom entry, enabling users to describe their condition without restrictions.

5.3 User Prediction History

The system maintains a prediction history feature that allows users to view their previous symptom inputs and corresponding prediction results. This helps users track their health-related queries over time and provides easy access to past reports.

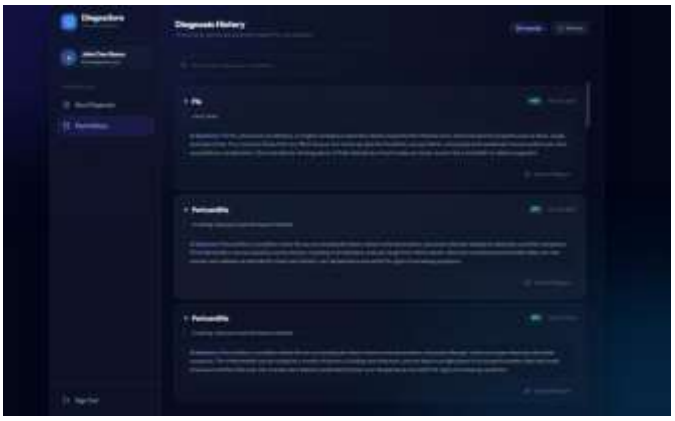

Fig. 5.3: User Prediction History

The history feature improves user experience by enabling quick reference to earlier predictions without re-entering data.

5.4 Output Prediction

After processing the input, the system generates disease predictions along with probability scores. In addition, the system provides recommended precautions

based on the predicted condition. The precautions guide users on initial steps to manage symptoms before consulting a doctor, improving the practical usability of the system. The prediction results are displayed clearly, ensuring easy interpretation by users. The system analyzes the severity of symptoms and categorizes cases into different urgency levels. This helps users prioritize their medical needs effectively.

Fig. 5.4: Disease Prediction with Precautions and Urgency

The system generates a detailed report that includes prediction results and explanations. Users can download this report for future reference or medical consultation. This feature enhances usability and supports real-world application.

5.5 Explainability using SHAP

The system provides explainable results using SHapley Additive exPlanations (SHAP), which highlight the contribution of each symptom to the prediction. This improves transparency and builds user trust in the system.

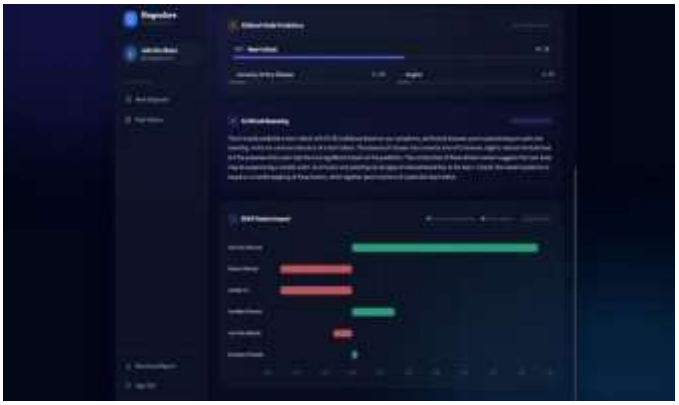

Fig. 5.5: Explainability using SHAP with LLM explanations

The system provides detailed explanations in natural language using large language models, making the results easy to understand.

5.6 Hospital and Doctor Recommendation

Based on the predicted condition, the system recommends relevant hospitals and doctors to the user. This helps users take appropriate action after receiving predictions.

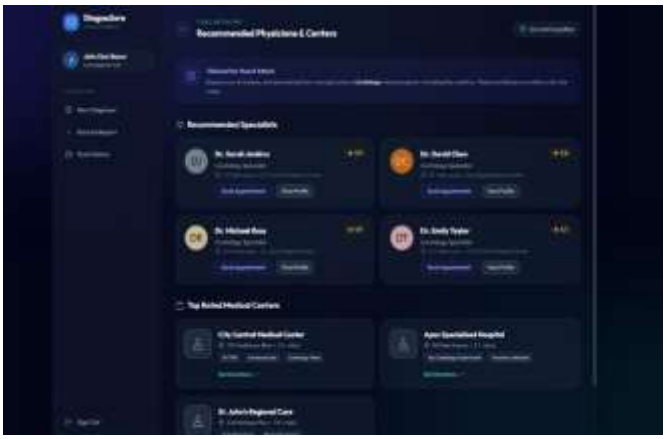

Fig. 5.6: Hospital and Doctor Recommendation

The system enables users to book appointments with recommended doctors directly. This feature reduces delays in seeking medical consultation.

5.7 Map Integration

The system integrates map services to display nearby healthcare facilities, improving accessibility.

Fig. 5.7: Map Integration

The map interface allows users to easily locate and navigate to hospitals.

6. CONCLUSION

DiagnoSure presents an integrated AI-powered healthcare support system that combines NLP, machine learning, and explainable AI to provide accurate and interpretable disease predictions. The use of BioBERT enables effective processing of unstructured symptom inputs, while XGBoost ensures high prediction accuracy. SHAP enhances transparency by explaining model decisions. Beyond prediction, the system incorporates practical healthcare features such as hospital and doctor recommendations, appointment booking, urgency detection, map integration, and LLM-based explanations. These features significantly improve usability and real-world applicability. The proposed system addresses key challenges in healthcare AI and serves as a reliable and scalable decision-support tool for early diagnosis and

medical assistance. The system demonstrates significant practical potential in integrating explainable AI into healthcare applications.

7. FUTURE SCOPE

The system can be extended to support multilingual inputs and incorporate medical imaging and laboratory data for more comprehensive diagnosis. Further improvements in AI models can enhance prediction accuracy and generalization across diverse medical conditions. Additionally, advancements in large language model-based explanations can improve personalization, user understanding, and interaction.

ACKNOWLEDGEMENT

We express our sincere gratitude to our project guide, Ms. Silja Varghese, Assistant Professor, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, for her invaluable guidance, encouragement, and continuous support throughout this work. Her expertise and insights significantly contributed to improving the quality of the project. We also thank the faculty members of the department for providing the necessary resources and academic support. Finally, we extend our heartfelt appreciation to our family, friends, and well-wishers for their constant motivation and encouragement.

REFERENCES

[1]. J. Lee, W. Yoon, S. Kim et al.: "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," Bioinformatics, 36(4), 1234-1240, 2020

[2]. F. Yi, H. Yang, D. Chen, Y. Qin, H. Han, J. Cui, W. Bai, Y. Ma, R. Zhang, H. Yu: "XGBoost-SHAP-based interpretable diagnostic framework for Alzheimer's disease," BMC Medical Informatics and Decision Making, 23, 137, 2023

[3]. H. Kim, et al.: "Refined XGBoost with SHAP Explainability for Non-Invasive Early Detection of Diabetic Kidney Disease: Estimated Cardiac Output as a Potential Predictor," Computers in Biology and Medicine, 145, 2025, pp. 105540

[4]. Myszewski, J. J., et al.: "BioBERT - RxReadmit: Improving Hospital Readmission Predictions Through Clinical Text Analysis with BioBERT," Frontiers in Digital Health, 4, 2025, pp. 878369

[5]. Hu, J., et al.: "Accurate Medical Named Entity Recognition Through Specialized NLP Models," Journal of Biomedical Informatics, 133, 2024, pp. 103473

[6]. Tarabanis, C., et al.: "Explainable SHAP-XGBoost Models for In-Hospital Mortality Prediction in Myocardial Infarction Patients," Journal of Clinical Medicine, 12(10), 2023, pp. 3201

[7]. M. Si, C. Zhang, Y. Chen et al.: "Explainable AI in Disease Prediction: A Systematic Literature Review," Journal of Biomedical Informatics, 129, 104073, 2022

[8]. C. Wang, J. Lee, K. Lin: "Medical Named Entity Recognition with Domain-Specific Pretrained Language Models," Journal of Biomedical Semantics, 12, 35, 2021

[9]. H. Rasmy, H. Wu, J. Wang, R. G. Xie: "MedBERT: Pretrained Contextualized Embeddings on Large-Scale Structured Electronic Health Records for Disease Prediction," NPJ Digital Medicine, 4, 86, 2021

[10]. Tomoo Inoue et al.: "XGBoost, a Machine Learning Method, Predicts Neurological Recovery in Patients with Cervical Spinal Cord Injury," Neurotrauma Reports. 2020 Jul 23;1(1):8–16.

[11]. X. Zhang, Q. Liu, R. Sun: "ClinicalBERT for ICD Coding: Automatic Disease Classification from Clinical Text," Journal of Biomedical Informatics, 107, 103482, 2020

[12]. Devlin, M. Chang, K. Lee, K. Toutanova: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019, pp. 4171–4186

[13]. J. Song, Y. Liu, R. Feng et al.: "ClinicalBERT: Modeling Clinical Notes and Predicting Patient Outcomes," arXiv preprint arXiv:1904.05342, 2019