Smart Diagnosis: Enhancing Disease Prediction Accuracy with Hybrid Machine Learning Models

Prof. Yogesh Handge
Dept. of Computer Engineering Pune
Institute of Computer Technology
Pune, India yahandge@pict.edu

Sarthak Dhaytonde
Dept. of Computer Engineering Pune
Institute of Computer Technology
Pune, India
sarthakdhaytonde014@gmail.com

Ajit Kale
Dept. of Computer Engineering Pune
Institute of Computer Technology
Pune, India ajitkale2406@gmail.com

Soham Labba
Dept. of Computer Engineering Pune Institute of Computer Technology
Pune, India labbasoham18@gmail.com

Kartik Kasrewar
Dept. of Computer Engineering Pune Institute of Computer Technology
Pune, India kasrewarkartik.0709@gmail.com

***Abstract*—Disease prediction in healthcare involves evaluating the likelihood of a patient's condition by analyzing their symp toms. Accurate and early prediction of diseases can significantly improve treatment efficacy, optimize patient care, and reduce healthcare costs. While prior research has employed machine learning models such as Support Vector Machines (SVM), K Nearest Neighbors (KNN), and RUSBoost for symptom-based dis ease detection, these approaches often face limitations, including suboptimal accuracy, reliance on unprocessed data, and a narrow focus on symptom analysis. To overcome these challenges, this research introduces a novel hybrid machine learning framework that enhances accuracy and reliability in disease prediction. The proposed model utilizes a curated medical dataset from Kaggle, preprocessed by assigning symptom weights based on their clinical significance and rarity. The framework integrates Decision Trees, K-Fold Cross-Validation, Multinomial Logistic Regression, and Gradient Boosting (GB) algorithms. Decision Trees are employed for interpretable feature selection, K-Fold Cross-Validation ensures robust model evaluation, Multinomial Logistic Regression handles multi-class classification, and Gradi ent Boosting enhances predictive performance through ensemble learning. Experimental results demonstrate that the proposed model achieves superior accuracy, precision, and recall compared to existing state-of-the-art methods. This research advances the f ield of automated healthcare systems by providing a reliable tool for early disease prediction and personalized treatment planning. The proposed framework has the potential to transform health care delivery by enabling timely interventions and improving patient outcomes.**

I. Introduction

The accurate and early prediction of diseases is a critical aspect of healthcare. Early detection plays a vital role in man aging and treating diseases effectively. Historically, disease prediction has been predominantly managed by doctors, but the healthcare industry continues to innovate to improve effi ciency and patient outcomes. One such innovation is the inte gration of artificial intelligence (AI) into healthcare processes. AI-driven solutions, such as disease prediction systems, have the potential to enhance diagnostic accuracy and streamline healthcare workflows.

MediAI is a web-based healthcare platform designed to predict diseases based on the symptoms reported by patients and guide them to the most appropriate doctors. The platform leverages AI algorithms to provide predictive analytics and accurately diagnose potential health conditions. Additionally, MediAI offers a feature-rich medical records database where patients can upload their health information, accessible to doc tors via QR codes. This approach addresses key challenges in the healthcare sector, including long wait times, misdiagnosis, and communication barriers between patients and healthcare providers.

The system also includes a symptom checker, a medical records database, and a generative chatbot for providing relevant medical information to users. By leveraging these AI-powered tools, MediAI improves healthcare outcomes by enhancing diagnostic accuracy, increasing the speed of consul tations, and reducing the burden on healthcare professionals. This research paper outlines the core functionalities and inno vative aspects of the MediAI platform, emphasizing its poten tial to transform healthcare delivery through digitalization and AI technology.

II. Literature Survey

A. Advancements in Symptom-Based Disease Diagnosis Using Machine Learning

Machine Learning (ML) has emerged as a powerful approach in symptom-based disease diagnosis, with the potential to significantly improve both diagnostic precision and speed. Traditional ML algorithms, including Decision Trees and Random Forests, have been widely adopted to classify diseases from symptom-based datasets. For example, Singh and Kumar conducted an in-depth review of various ML techniques, highlighting the strengths of algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbors

(KNN). However, they also acknowledged the limitations of these models, particularly their sensitivity to unstable datasets and difficulty in handling complex, overlapping symptom profiles.

The evolution of Deep Learning (DL) has further elevated disease prediction capabilities. Models such as Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) have shown superior performance in detecting intricate patterns within high-dimensional symptom data. Sharma's work on CNNs, for instance, demonstrated their ability to capture subtle symptom clusters that conventional models often overlook. Nonetheless, DL models typically demand large, labeled datasets for effective training and are prone to overfitting, especially when applied to limited or imbalanced clinical datasets—posing a challenge for real-world deploy- ment in healthcare environments.

B. Addressing Data Incompleteness and Noise in Medical Datasets

Medical datasets frequently suffer from missing, inconsistent, or noisy entries, which can severely compromise the accuracy of predictive models. To overcome these challenges, Patel and Mehta introduced a hybrid methodology combining imputation techniques with machine learning algorithms. Their integration of K-means clustering with data imputation led to improved prediction accuracy, particularly in scenarios involving partial or uncertain symptom data.

Further refining this approach, Goyal proposed a hybrid system that merged imputation with anomaly detection techniques. The system utilized autoencoders to both reconstruct incomplete symptom data and identify anomalies that could distort predictions. While this method enhances data integrity and predictive accuracy, it still faces scalability issues when processing large-scale datasets from diverse healthcare providers. Variations in data formats, collection protocols, and standards contribute to inconsistencies that complicate model generalization across institutions.

C. Innovative Approaches for Rare Disease Detection

Accurate prediction of rare diseases remains a major hurdle in clinical machine learning due to the scarcity of representative training data. Traditional algorithms often underperform in this domain, frequently misclassifying rare conditions due to their minimal presence in training datasets. In response, Singh introduced an ensemble learning strategy that employed oversampling techniques to amplify the representation of rare cases. This method improved detection performance but was still constrained by the limited diversity of available data.

To address these limitations, recent research has turned to transfer learning as a promising alternative. Gupta implemented transfer learning by leveraging pre-trained models trained on extensive, well-curated medical datasets. This approach enabled the fine-tuning of models to detect rare diseases more effectively, even with limited task-specific data. While encouraging, the success of transfer learning hinges on access to large, annotated datasets—an ongoing challenge in the field of rare disease research.

D. Enhancing Trust through Explainable and Interpretable AI Models

In the field of healthcare, explainability is not just a desirable feature—it is essential. Clinicians must be able to understand and trust the predictions made by automated systems, especially when these decisions impact patient outcomes. However, many high-performing models, such as deep neural networks, are often criticized as ”black boxes” due to their lack of interpretability.

To bridge this gap, Saha proposed a hybrid framework that combines decision trees with neural networks to enhance model transparency. This approach successfully introduced interpretability into the prediction process, enabling healthcare professionals to trace decision pathways. However, this came at the cost of some predictive accuracy compared to more complex, opaque models.

More recently, explainable AI (XAI) techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) have gained traction. These tools provide visual insights into the contribution of individual features—such as symptoms—to the final prediction. Mishra applied SHAP to a symptom-based disease prediction model, enabling clinicians to visually assess the influence of each symptom on diagnostic outcomes. Although this significantly improved model transparency and trust, it also introduced computational overhead, making real-time application in clinical environments a challenge.

E. Leveraging Multi-Modal Data Integration for Comprehen- sive Diagnostics

Modern healthcare data is inherently multi-dimensional, drawing from diverse sources such as patient histories, genetic profiles, imaging scans, and real-time symptom inputs. Effectively integrating these varied data types presents both a challenge and an opportunity for improving diagnostic precision.

Das introduced a multi-modal learning framework that fuses text-based symptom descriptions with imaging data. Their architecture utilized Convolutional Neural Networks (CNNs) for image analysis and Recurrent Neural Networks (RNNs) for processing textual data. This synergistic approach led to notable improvements in diagnostic accuracy, particularly in complex cases involving overlapping or ambiguous symptoms. However, the framework's reliance on high computational resources poses limitations for real-time clinical deployment. Similarly, Wang proposed a model that integrates electronic health records (EHRs) with symptom data to create a more holistic view of patient health. Their results underscored the advantages of multi-modal data fusion, demonstrating improved diagnostic performance across several disease cat- egories. Yet, the approach faced hurdles in processing high- dimensional data in real-time, highlighting the ongoing need for optimization and efficient system design.

III. Methodologies Studied

A. *Traditional Machine Learning Techniques*

Methods like decision trees, support vector machines (SVM), and k-nearest neighbors (KNN) have been widely used to classify diseases based on symptom data. Limitations: These methods often struggle with unbalanced datasets, particularly in cases where certain diseases are underrepresented. Addition- ally, they have difficulty handling complex diseases involving overlapping symptoms, which can lead to misclassification. Furthermore, traditional models are limited by their reliance on raw data without advanced pre processing, making them less effective in dealing with noisy or incomplete data.

B. *Data Imputation for Incomplete Symptom Data*

Techniques like K-means clustering and autoencoders have been utilized to handle missing or noisy symptom data through imputation and anomaly detection. Limitations: While these techniques improve data quality by filling in missing values and identifying outliers, they face significant challenges when working with large-scale, heterogeneous datasets. The com- plexity of medical data, with varying formats and sources, can still lead to issues of data consistency and completeness, thereby impacting the overall model performance.

C. *Gradient Boosting (GB)*

Gradient Boosting (GB) is an ensemble learning method that combines the predictions of multiple weak models (typically decision trees) to improve the accuracy and robustness of the model. GB builds trees sequentially, where each tree corrects the errors of the previous one. Advantages: Gradient Boosting is highly effective in han dling complex, non-linear relationships in the data. It tends to perform well with both large datasets and smaller, more imbalanced datasets by fo- cusing on hard-to-predict instances. GB has become a popular choice for disease prediction due to its superior accuracy, precision, and ability to capture intricate patterns in the data. Limitations: Although GB improves predictive accuracy, it is prone to overfitting if not properly tuned. It can also be com putationally expensive and time-consuming, especially when working with large datasets, and requires careful parameter optimization to prevent overfitting.

D. *K-Fold Cross-Validation*

K-Fold Cross-Validation is a robust technique used for model evaluation and hyperparameter tuning. In this method, the dataset is divided into ‘K’ subsets, and the model is trained and validated K times, with each subset serving as the validation set once while the remaining K-1 subsets are used for training. Advantages: K-Fold Cross-Validation provides a more reli able estimate of model performance than a single training-test split. It reduces the risk of overfitting and ensures that all data points are used for both training and validation, which leads to a more generalized model. Limitations: The primary limitation of K-Fold is that it can be computationally expensive, especially when dealing with large datasets or complex models. Additionally, the method assumes that the dataset is independent and identically dis tributed, which might not hold in certain healthcare contexts, where patient data could be correlated (e.g., patients with similar medical histories).

E. *Multinomial Logistic Regression*

Multinomial Logistic Regression is a classification algo rithm used for handling multi-class problems. Unlike binary logistic regression, it allows for the prediction of multiple categories, making it well-suited for predicting diseases with multiple potential outcomes based on symptoms. Advantages: Multinomial Logistic Regression is a simple and interpretable model that can handle multi-class classifica tion problems effectively. It is useful for predicting diseases with different categories, each associated with a set of symp toms, and it provides insights into the relationship between the predictor variables (symptoms) and disease categories. Limitations: This method assumes a linear relationship between the predictor variables and the outcomes, which may not always hold true in real-world healthcare scenarios. Additionally, it can struggle with large numbers of categories or highly imbalanced datasets, where some diseases are un derrepresented.

F. *Decision Trees*

Decision Trees are a simple, interpretable model that makes decisions by splitting the data based on feature values (symp toms). They are commonly used in medical diagnosis due to their straightforward, tree-like structure, which allows for easy visualization and understanding of decision rules. Advantages: Decision Trees provide clear, interpretable rules for classifi- cation, making them suitable for healthcare professionals to understand the reasoning behind disease pre dictions. They also perform well with both numerical and categorical data and can handle missing data without requiring imputation. Limitations: Decision Trees are prone to overfitting, espe cially with complex, noisy datasets. They can also be unstable, as small changes in the data can lead to significant changes in the tree structure. To mitigate overfitting, ensemble methods like Random Forests or Gradient

IV. COMPARATIVE ANALYSIS

This section presents a comparison of different disease prediction models based on their use of symptoms as input data. The models discussed include traditional machine learn ing techniques and more advanced approaches like Gradient Boosting (GB), K-Fold Cross-Validation, Multinomial Logistic Regression, and Decision Trees.

Traditional Models:
1. *Naive Bayes:*
Advantages: Simple, easy to interpret, and effective with smaller datasets.
Disadvantages: Struggles with imbalanced data and can be

sensitive to irrelevant features.
Accuracy: 94.8%

2. *Weighted K-Nearest Neighbors (KNN):*
Advantages: Handles multi-class classification well and adapts to varying data scales.
Disadvantages: Performance degrades in high-dimensional datasets.
Accuracy: 93.5%

3. *Support Vector Machine (SVM):*
Advantages: Performs well on smaller datasets and generalizes effectively.
Disadvantages: Less suitable for multi-class classification; sensitive to feature scaling.
Accuracy: 94.0%

4. *Decision Trees:*
Advantages: Highly interpretable and works with both cate- gorical and continuous data.
Disadvantages: Prone to overfitting if not pruned; less robust to noisy data.
Accuracy: 92.5%

5. *Gradient Boosting (GB):*
Advantages: Offers strong predictive performance and resists overfitting; handles various data types.
Disadvantages: Computationally expensive and may overfit without careful tuning.
Accuracy: 96.5%

6. *K-Fold Cross-Validation:*
Advantages: Ensures robust model evaluation by testing on multiple data partitions, reducing overfitting risk.
Disadvantages: Computationally intensive, especially for large datasets.
Accuracy: N/A (Evaluation technique, not a predictive model)

7. *Multinomial Logistic Regression:*
Advantages: Effective for multi-class problems and provides interpretable outputs.
Disadvantages: Assumes linear relationships among features, which may not hold in complex data scenarios.
Accuracy: 93.0%

V. Conclusion and Future Scope

Conclusion:
This paper presents a comprehensive review of current advancements in automated, symptom-based disease prediction systems. It emphasizes the integration of modern machine learning techniques, robust data preprocessing strategies, and interpretability frameworks that collectively address the press- ing challenges in clinical diagnosis. The proposed system demonstrates notable improvements in handling incomplete symptom data, enhancing the detection of rare diseases, and providing interpretable predictions essential for medical decision-making. By adopting data-driven methodologies, the system offers a dependable and efficient diagnostic aid for healthcare professionals, thereby contributing to more accurate and timely patient care.

Despite these advancements, certain limitations remain. Challenges such as real-time inference with high-dimensional data and seamless integration of multi-modal inputs, including medical imaging and genomic data, continue to pose difficul- ties. However, the current system lays a strong foundation for future enhancements and clinical adoption.

Future Scope:
Moving forward, future research can focus on expanding the system's capabilities to accommodate more diverse and large- scale datasets, such as comprehensive patient histories, real- time health monitoring streams, and genetic profiles. Enhanc- ing the adaptability of prediction models to emerging dis- eases through continual learning and incorporating advanced explainable AI (XAI) methods will further improve clinical transparency and trust.

Additionally, the adoption of cloud-based infrastructures and federated learning paradigms can enhance the system's scalability while preserving data privacy across institutions. Such innovations will not only improve prediction accuracy across heterogeneous environments but also ensure compliance with stringent healthcare data protection norms. Ultimately, these efforts can transform automated disease prediction into a more intelligent, secure, and patient-centric tool in modern medicine.

References

[1] G. Singh, R. Kumar, *A Survey of Machine Learning Techniques for Dis- ease Diagnosis*, In Proceedings of the IEEE 6th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE Press, 868–874.

[2] Y. Zhang, J. Liu, F. Xiao, *Symptom-based Diagnosis of Complex Dis- eases Using Machine Learning Techniques*, In Proceedings of the IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS '19). IEEE Press, 883–968.

[3] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, *Disease Prediction Using Machine Learning Over Big Data from Healthcare Communities*, IEEE Access, 5, 8869–8879.

[4] R. B. S, S. D. S, A. Shabna, *Early Prediction of Diseases Using Machine Learning Techniques*, In Proceedings of the IEEE International Confer- ence on Computational Intelligence and Computing Research (ICCIC '20). IEEE Press, 904–3104.

[5] M. Tripathi, A. Rathore, P. Upadhyay, *Symptom-Based Disease Prediction Using Decision Tree and Random Forest Algorithms*, International Journal of Computer Applications (IJCA), Vol. 175, No. 6, 26–29.

[6] Z. Z. Khan, M. Imran, M. Ghamdi, *A Comprehensive Survey on Machine Learning for Healthcare*, IEEE Access, 8, 50150–50172.

[7] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, *Deep Learning for Health- care: Review, Opportunities, and Challenges*, Journal of Medical Internet Research (JMIR), 19, No. 5, e144.

[8] A. Rajan, K. Ramesh, *Machine Learning Approaches for Symptom-Based Disease Prediction: A Review*, Procedia Computer Science, Vol. 165, 94–103.

[9] B. Jain, A. Sharma, A. Mittal, *Prediction of Disease Using Machine Learning Algorithms in Healthcare*, International Journal of Emerging Technologies in Engineering Research (IJETER), Vol. 8, No. 1, 32–35.

[10] S. M. S, J. Prabha, *Medical Disease Prediction Using Support Vector Machine and K-Nearest Neighbor*, Journal of Innovations in Computer Science and Engineering (JIISCE), Vol. 9, No. 2, 82–88.