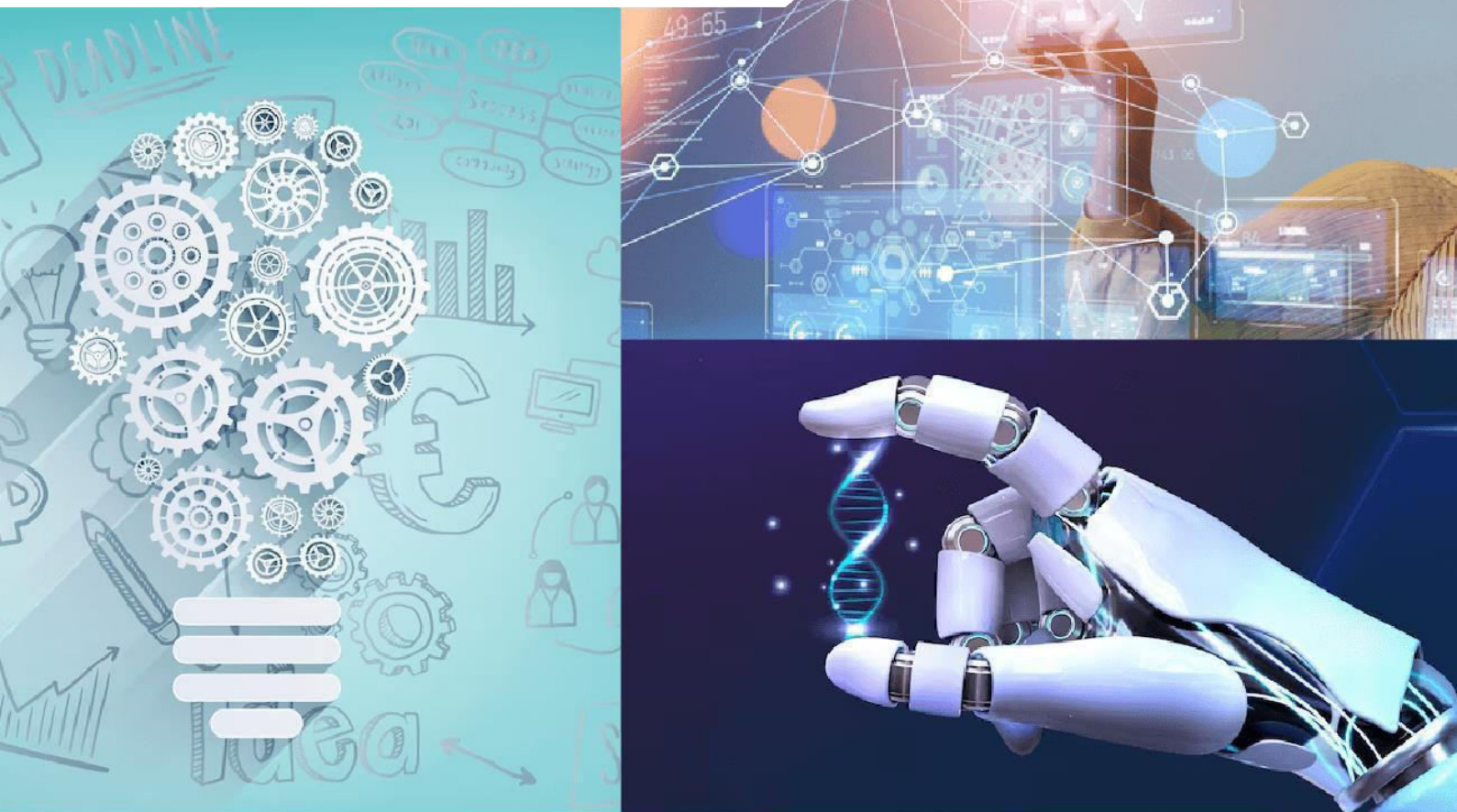




International Journal of Advanced Research in Education and Technology (IJARETY)

Volume 13, Issue 3, May - June 2026

Impact Factor: 9.971



AI Model for Predicting Multiple Diseases Using Ensemble Techniques

Hirpara Sachin Vinubhai, Prof. Rajesh N

Student, Department of MCA, AMC Engineering College, Bengaluru, India

Assistant Professor, Department of MCA, AMC Engineering College, Bengaluru, India

ABSTRACT: I design of medical diagnosis systems has undergone a dramatic change due to the quick expansion of healthcare data and the development of artificial intelligence. Early and precise disease detection is crucial for enhancing patient outcomes and lowering death rates in contemporary healthcare settings. When applied to complicated and real-world medical datasets traditional machine learning techniques frequently rely on single predictive models which may have drawbacks like high variance bias and poor generalization. This study suggests an AI-based multi-disease prediction system that makes use of ensemble learning techniques to improve predictive accuracy robustness and reliability in order to address these issues. Using stacking and voting techniques the suggested system unifies several machine learning algorithms including Random Forest Support Vector Machine Logistic Regression and Gradient Boosting into a single ensemble framework. Through the analysis of clinical and pathological datasets the model is intended to predict various diseases such as diabetes heart disease and liver disorders. In order to improve the quality of the data extensive experiments were carried out using publicly available healthcare datasets. Appropriate preprocessing techniques including missing value handling feature scaling feature selection and data balancing were used. Standard metrics such as Accuracy Precision Recall F1-Score and ROC-AUC were used to assess the systems performance. The results unequivocally show that the ensemble model outperforms individual classifiers in terms of overall robustness and prediction stability. Terms in Index. AI in healthcare machine learning ensemble learning multi-disease prediction predictive analytics and clinical decision support systems.

KEYWORDS: Disease Prediction, Machine Learning, Ensemble Learning, Healthcare AI, Multi-Disease Prediction

I. INTRODUCTION

Making timely and accurate decisions is essential to saving lives in the healthcare industry which is one of the most important fields. Artificial intelligence has become a potent tool to help medical professionals with diagnosis and prediction tasks due to the quick digitization of medical records and the growing accessibility of large-scale healthcare datasets.

Disease diagnosis has traditionally been largely dependent on the experience and knowledge of physicians which while extremely valuable can occasionally be impacted by human error workload pressure and the complexity of patient conditions. In order to automate disease prediction systems machine learning techniques have been widely used in recent years. These techniques analyze past patient data and find hidden patterns that are difficult to see through manual inspection. Nevertheless the complexity and variability found in medical datasets are too much for the majority of conventional machine learning models such as Logistic Regression Support Vector Machine or Decision Trees to handle on their own.

Overfitting underfitting and sensitivity to noisy data are common problems for these models which eventually impacts prediction accuracy. In the field of healthcare analytics ensemble learning techniques have become increasingly important in order to overcome these limitations. Similar to how medical professionals weigh several viewpoints before making a final diagnosis ensemble learning operates on the premise of combining several base models to produce a more accurate and stable final prediction.

In order to make it more useful and effective for actual healthcare applications an ensemble-based AI system for predicting several diseases within a single framework is suggested in this study.

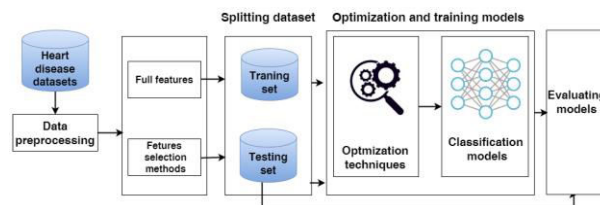


Fig. 1. Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction

II. RELATED WORK OR LITERATURE REVIEW

The application of machine learning methods for disease prediction in healthcare systems has been investigated by a number of researchers. Early research mostly concentrated on using individual classifiers such as Decision Trees Naïve Bayes and Logistic Regression to predict conditions like diabetes and heart disease.

When applied to larger and more complex medical data these models frequently failed to generalize well despite offering respectable accuracy on small datasets. Due to their capacity to manage non-linear relationships and lessen overfitting Support Vector Machines and Random Forest algorithms were later developed by researchers which increased classification accuracy. Nevertheless these models continued to function as stand-alone systems and had limited capacity to handle scenarios involving the prediction of multiple diseases. Ensemble learning techniques which combine several machine learning models to enhance overall performance have become more popular in more recent research. By minimizing the shortcomings of individual models research has demonstrated that stacking and voting-based ensemble techniques greatly improve prediction accuracy in healthcare applications.

Despite these developments there is little research on unified frameworks that can handle several diseases at once using a single integrated model and the majority of current systems are built to predict a single disease at a time. The development of the suggested system is based on this gap in the body of current literature. Despite these developments the majority of current systems are made for single-disease prediction concentrating on a single illness at a time. Research on unified frameworks that can use a single integrated system to predict several diseases at once is scarce. Furthermore when applied to various healthcare datasets many of the current models lack adequate scalability and real-world adaptability.

Additionally some research has not adequately addressed significant issues like feature selection data imbalance and model interpretability in clinical settings. These drawbacks emphasize the requirement for more reliable and adaptable systems.

Therefore, creating a multi-disease prediction system using an ensemble learning approach that can effectively handle multiple diseases while maintaining high accuracy and reliability is the gap in the body of current literature. The suggested system in this study is motivated by this.

III. PROPOSED SYSTEM

The suggested system is built as a full machine learning pipeline that begins with gathering data and concludes with making a final disease prediction. Numerous publicly accessible healthcare datasets including those on diabetes heart disease and liver disease are used by the system.

Important clinical characteristics like age blood pressure glucose cholesterol bilirubin and other medical parameters are included in these datasets. To prevent bias in classification results this involves balancing datasets using oversampling techniques like SMOTE normalizing numerical features using standard scaling techniques handling missing values using statistical methods and encoding categorical variables into numerical format.

Following preprocessing feature selection is carried out to find the most pertinent characteristics that have a major impact on disease prediction. Correlation analysis and feature importance scores produced by tree-based models like Random Forest are used for this. Several machine learning models such as Logistic Regression Support Vector Machine Random Forest and Gradient Boosting are independently trained as base learners after the dataset has been optimized.

By capturing distinct patterns in the data each model contributes to an increase in the overall diversity of predictions. For effective processing a layered architecture was used in the systems construction.

These are the main layers.

- Data Collection Layer: gathers clinical parameters lab reports and medical records from healthcare datasets.
 - Data Preprocessing Layer: improves data quality by handling missing values encoding normalization and feature selection.
 - Model Training Layer: trains a variety of machine learning models including Random Forest Logistic Regression Support Vector Machine and Gradient Boosting.
 - Ensemble Layer: creates a final output by combining predictions using voting and stacking techniques.
 - Prediction Layer: gives a final disease prediction based on patient data input.
 - Visualization Layer: Dashboards are used to display results for improved decision-making and interpretation.
- The system is scalable modular and appropriate for practical healthcare applications because of this architecture.

IV. METHODOLOGY

In order to predict diseases the suggested system uses an organized machine learning pipeline. A thorough preprocessing step is used to enhance data quality because raw medical data frequently contains missing values inconsistencies and noise.

A. Data Preprocessing

Preprocessing includes:

- using statistical techniques to handle missing values.
- standard normalization methods for feature scaling.
- putting categorical variables into numerical form through encoding.
- To prevent classification bias use SMOTE for data balancing.

B. Feature Selection

Following preprocessing the most pertinent characteristics that greatly influence disease prediction are found through feature selection. This is accomplished through the use of feature importance scores produced by tree-based models like Random Forest and correlation analysis.

C. Model Training

Following dataset optimization several machine learning models are trained separately as base learners. They consist of the following.

- Logistic Regression
- Support Vector Machine
- Random Forest
- Gradient Boosting

Each model captures different patterns in the data, improving diversity and prediction strength.

D. Ensemble Learning

The trained base models are combined using ensemble learning techniques:

- Voting Classifier: Combines predictions based on majority or weighted voting.
- Stacking Classifier: Uses a meta-learner trained on base model outputs to generate final predictions.

This ensemble approach reduces individual model errors and improves generalization capability. Preprocessing includes:

- Missing value handling
- Feature scaling
- Encoding categorical variables
- Data balancing using SMOTE Models used:
- Logistic Regression
- Support Vector Machine
- Random Forest
- Gradient Boosting Ensemble methods:
- Voting Classifier
- Stacking Classifier

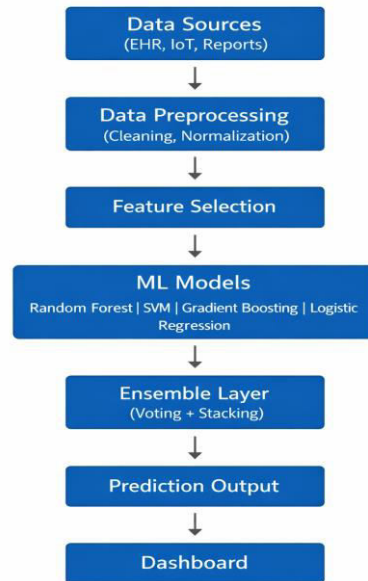


Fig. 2. Proposed System Architecture for Multi-Disease Prediction

V. MATHEMATICAL MODEL

Let the healthcare dataset be represented as:

$$D = \{x_1, x_2, x_3, \dots, x_n\}$$

where D denotes the complete dataset and x_i represents each patient record consisting of clinical attributes such as age, blood pressure, glucose level, cholesterol, and other medical parameters.

For multi-disease prediction, each patient record is mapped to a label set:

$$Y = \{y_1, y_2, \dots, y_m\}$$

where m represents the number of diseases (e.g., diabetes, heart disease, liver disease), and $y_j \in \{0, 1\}$ indicates absence or presence of disease.

Each machine learning model is trained on dataset D and produces a predictive function:

$$M_i = f_i(D)$$

where f_i is the learning function of the i^{th} model. Each model outputs a probability score:

$$P_i(y_j = 1|x) = f_i(x)$$

where P_i represents the probability of disease j for input patient x .

The final ensemble prediction is computed using averaging:

$$P(y_j = 1|x) = \frac{1}{k} \sum_{i=1}^k P_i(y_j = 1|x)$$

For weighted ensemble learning:

$$P(y_j = 1|x) = \sum_{i=1}^k w_i P_i(y_j = 1|x)$$

where:

$$\begin{aligned} & k \\ & X \\ & w_i = 1 \\ & i=1 \end{aligned}$$

Decision Rule (Final Prediction)

The final classification decision is:

$$y_j = \{1, \text{if } P(y_j = 1|x) \geq 0.50, \text{otherwise}\}$$

Performance Metrics

To evaluate model performance, standard metrics are used: Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

$$Recall = \frac{TP}{TP + FN}$$

F1-Score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Loss Function (for Optimization)

For probabilistic classification, cross-entropy loss is used:

$$L = - \sum_{j=1}^m [y_j \log(p_j) + (1 - y_j) \log(1 - p_j)]$$

where p_j is predicted probability.

The ensemble model shows higher accuracy and stability compared to individual models due to reduced variance and improved generalization capability.

Model Comparison

TABLE I
MODEL COMPARISON

Model	Accuracy	F1 Score	Stability
SVM	High	Medium	Medium
RF	High	High	High
GB	Very High	High	High
Ensemble	Very High	Very High	Very High

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The suggested systems performance was assessed using common machine learning evaluation metrics including ROCAUC F1-Score Accuracy Precision and Recall. According to experimental results Random Forest and Gradient Boosting perform better because they can handle complex patterns while individual models like Logistic Regression and SVM achieve moderate accuracy.

Nonetheless the suggested ensemble model consistently achieves greater accuracy and improved stability across various datasets outperforming all individual classifiers. The combination of several models which lowers bias and variance and increases overall prediction reliability is responsible for the performance improvement. Because it can learn from base model outputs more successfully than the voting method the stacking-based ensemble method performs marginally better. Overall the findings show that medical diagnosis tasks involving several diseases benefit greatly from ensemble learning.

Combining several models improves overall prediction reliability and lowers bias and variance which is responsible for the performance improvement. Because it can learn from base model outputs more efficiently the stacking-based ensemble method outperforms the voting method by a small margin. Overall the findings demonstrate the great efficacy of ensemble learning in medical diagnosis tasks involving several diseases.

Additionally even when using unseen or slightly unbalanced datasets the ensemble approach yields more stable predictions. Additionally it increases robustness by reducing the influence of outliers and noisy data. The suggested systems dependability is demonstrated by its consistent performance across various assessment metrics. These findings demonstrate the models applicability to actual healthcare settings.

VII. APPLICATIONS, ADVANTAGES, AND LIMITATIONS

A. Advantages

By combining several machine learning models including Random Forest Support Vector Machine Logistic Regression and Gradient Boosting the suggested system attains high prediction accuracy. By reducing the shortcomings of individual models this ensemble approach enhances overall performance. False positives and false negatives which are crucial in healthcare applications are greatly decreased. As a result disease prediction results are more dependable and credible. Additionally early disease detection is supported by the system which lowers the risk of serious medical conditions and aids in preventive healthcare.

B. Limitations

High-quality datasets are essential to the systems operation. Prediction accuracy may be adversely affected if the dataset has imbalance noise or missing values.

Due to the training and combination of several machine learning models which lengthens processing times and uses more memory it has a high computational complexity.

Low-power devices or real-time systems without optimization are less suited for the training process because it requires substantial computational resources.

Particularly in ensemble models like Random Forest and Gradient Boosting the systems interpretability is limited making it challenging to clearly explain predictions to nontechnical users.

As medical patterns and datasets change over time accuracy must be maintained through ongoing model retraining and updates.

If appropriate preprocessing methods are not used performance may deteriorate when applied to entirely new or highly unbalanced datasets.

VIII. IMPLEMENTATION

Data collection preprocessing model training ensemble learning and prediction are all part of the full machine learning pipeline used in the implementation of the suggested AIbased multi-disease prediction system. Python and libraries like Scikit-learn Pandas NumPy and Matplotlib are used in the systems development.

Data collection is the first step in the process and datasets related to diabetes heart disease and liver disease are used. Clinical characteristics like age blood pressure glucose level cholesterol and other medical parameters are included in these datasets. During the data preprocessing phase statistical techniques are used to handle missing values categorical variables are encoded and feature scaling is implemented.

The dataset is balanced and model performance is enhanced using SMOTE. Correlation analysis and Random Forest feature importance are then used for feature selection in order to eliminate unnecessary attributes and boost productivity. Several machine learning models including Random Forest

Logistic Regression Support Vector Machine and Gradient Boosting are trained as base learners during the model training phase.

Following training all models predictions are combined using ensemble learning techniques like voting and stacking which increase accuracy and decrease errors. Lastly to aid in improved interpretation and decision-making the system uses visualization tools to display the results of disease prediction on fresh patient data.

The system is scalable modular and appropriate for use in practical healthcare settings.

IX. FUTURE WORK

This system can be improved in the future with an emphasis on increasing accuracy scalability and practical usability. Integrating deep learning models like neural networks to identify more intricate patterns in medical data and enhance prediction performance beyond conventional machine learning techniques is one crucial path.

Additionally real-time healthcare data processing can be added to the system enabling ongoing patient health monitoring via wearable technology and hospital information systems. In actual clinical settings this would make the model more useful. Federated learning is another significant improvement that improves privacy and security in healthcare applications by allowing model training across several hospitals without exchanging sensitive patient data.

Additionally by using explainable AI (XAI) techniques to improve model interpretability physicians will be better able to comprehend how predictions are made which will boost system trust.

X. CONCLUSION

In summary this study introduces an ensemble learningbased AI-based multi-disease prediction system that greatly increases the precision and dependability of medical diagnosis. The system successfully overcomes the limitations of individual classifiers by combining several machine learning models including Logistic Regression SVM Random Forest and Gradient Boosting into a single ensemble framework.

The experimental findings show that the suggested model outperforms standalone models indicating its suitability for practical healthcare applications. The system offers a robust and scalable framework for multi-disease detection in addition to improving prediction accuracy. This system has the potential to become an advanced intelligent healthcare assistant for physicians and other medical professionals with additional developments like federated learning deep learning integration and real-time data processing.

Additionally by offering prompt and data-driven diagnostic support the suggested strategy can greatly lessen the workload of medical professionals.

REFERENCES

- [1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [2] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, pp. 273–297, 1995.
- [3] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832–844, 1998.
- [4] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
- [5] K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
- [6] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann, 2011.
- [7] A. Kumar and R. Sharma, "Predictive Analytics in Healthcare Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 182–194, 2020.
- [8] S. Zhang, X. Li, and M. Zong, "Ensemble Learning Methods for Medical Diagnosis: A Review," Artificial Intelligence in Medicine, vol. 102, 2020.
- [9] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.
- [10] World Health Organization (WHO), "Global Health Observatory Data Repository," 2023. Available: <https://www.who.int>
- [11] Kaggle Datasets, "Healthcare Dataset Repository for Machine Learning," 2024. Available: <https://www.kaggle.com>
- [12] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," Journal of Physics: Conference Series, vol. 1142, 2018.



International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 9.971