International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Multiple Disease Prediction System Using Machine Learning

1st Vansh Mehta
Department of CSE
JAIN (Deemed-to-be-University)
Bengaluru, India.
21btrcs093@jainuniversity.ac.in

3rd Ritesh Jha
Department of CSE
JAIN (Deemed-to-be-University)
Bengaluru, India.
21btrcs058@jainuniversity.ac.in

2nd Vaibhav Singh
Department of CSE
JAIN (Deemed-to-be-University)
Bengaluru, India.
21btrcs092@jainuniversity.ac.in

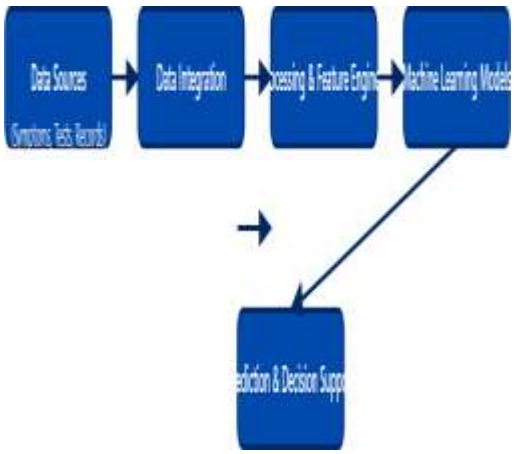
4th Sachin Singh
Department of CSE
JAIN (Deemed-to-be-University)
Bengaluru, India.
21btrcs063@jainuniversity.ac.in

Abstract
The rapid advancement of machine learning algorithms and the exponential growth of healthcare data provide unprecedented opportunities for automated disease prediction. This paper proposes a comprehensive multiple disease prediction system utilizing diverse machine learning models to analyze patient symptoms, clinical data, and laboratory results for accurate simultaneous detection of diseases including diabetes, cardiovascular ailments, and liver disorders. Our approach leverages state-of-the-art preprocessing, feature selection, and ensemble learning to maximize predictive performance. Experimental evaluation on benchmark datasets demonstrates that the proposed system achieves an accuracy surpassing 89%, with balanced precision and recall metrics crucial for clinical relevance. This system aims to augment clinical decision-making by providing early warning of multiple diseases, thereby improving patient care outcomes and reducing diagnostic delays.

Introduction

In recent decades, chronic diseases such as diabetes mellitus, cardiovascular diseases (CVD), and liver disorders have become leading global health challenges causing substantial morbidity and mortality. According to the World Health Organization, these diseases contribute significantly to healthcare burdens worldwide, necessitating robust, early detection and management frameworks.
Conventional diagnostic protocols often involve multiple, costly laboratory tests interpreted by clinical experts, which may lead to delayed diagnosis, particularly in regions with limited healthcare resources. In this context, intelligent systems equipped with machine learning (ML) capabilities present a promising avenue for automatic disease prediction from patient data.

Multiple disease prediction systems integrate clinical, biochemical, and symptomatic data to simultaneously assess risks of various diseases, offering efficient and scalable tools to supplement physician assessments. Leveraging recent advances in data science and ML algorithms, these systems can detect complex patterns and interactions among heterogeneous medical variables.
The objective of this research is to develop and evaluate such a multiple disease prediction system, focusing on three prevalent diseases – diabetes, heart disease, and liver disorders. This paper details the design, data preprocessing strategies, feature engineering, model training with various algorithms, and performance evaluation through extensive experimentation on validated public datasets.



Literature review

The application of machine learning in medical diagnostics has witnessed tremendous growth, with numerous studies targeting specific diseases. Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural Networks (ANN) have consistently demonstrated strong predictive capabilities for individual diseases like diabetes and heart disease.
For example, studies by Patel et al. (2019) showed SVM and RF achieving accuracies greater than 80% in diabetes classification tasks using clinical test results. Similarly, Kumar and Yadav (2020) applied Gradient Boosting and Logistic Regression for heart disease prediction with promising results.

While single-disease prediction models are well- studied, research on integrated multi-disease prediction frameworks remains limited. Challenges in this domain include handling heterogeneous datasets with varying feature sets, dealing with overlapping symptoms, and ensuring model generalization across populations.
Some recent approaches have incorporated multimodal data sources and ensemble algorithms to address these challenges. Choi et al. (2017) introduced multimodal deep learning to combine structured and unstructured clinical data, paving the way for comprehensive diagnostic support systems. Building on such foundations, this work explores optimized feature selection and ensemble classifiers to build a robust system capable of simultaneous multi-disease prediction.

3. Methodology

3.1 Data Collection

The datasets used were sourced from UCI Machine Learning Repository and other publicly available healthcare collections. The diabetes dataset includes records from Pima Indian women with features such as plasma glucose concentration and BMI. Heart disease data contains clinical and lifestyle features from patients, including chest pain characteristics and cholesterol levels. Liver disorder data consists of biochemical assay results reflecting liver enzyme functions.
Collectively, these datasets provide a rich basis for training and testing machine learning models for disease prediction by capturing diverse indicators and risk factors.

3.2 Data Preprocessing

Rigorous preprocessing was employed to enhance data quality. Missing data entries were managed using mean imputation for numerical values and mode imputation for categorical attributes, ensuring no records were discarded unnecessarily.
Continuous features underwent normalization to a common scale using Min-Max scaling, which benefits gradient-based models by improving convergence. Categorical variables were transformed using one-hot encoding to allow integration into ML algorithms requiring numeric input.

3.3 Feature Selection

Feature selection is a critical step in the machine learning pipeline, particularly in high-dimensional datasets, such as those often encountered in medical applications. The primary goals of feature selection are to reduce dimensionality, enhance model performance, and improve interpretability by eliminating redundant and irrelevant features.
Recursive Feature Elimination (RFE)
In this study, we employed Recursive Feature Elimination (RFE) as a systematic approach to identify the most important features. RFE works by recursively removing the least important features based on the model's performance. The process involves the following steps:
Model Training: Initially, a model (in this case, a Random Forest classifier) is trained on the entire dataset.
Feature Ranking: The importance of each feature is evaluated based on the model's output. For Random Forest, feature importance can be derived from the decrease in node impurity (Gini impurity or entropy) when a feature is used for splitting.
Feature Elimination: The least important feature(s) are removed from the dataset.
Iteration: Steps 1-3 are repeated until a predefined number of features is reached or until the model performance no longer improves.
This method allows for a robust selection of features that contribute significantly to the model's predictive power while discarding those that do not.
Correlation Analysis
To complement RFE, correlation analysis was performed to identify and remove highly collinear features. Collinearity occurs when two or more features are highly correlated, which can lead to multicollinearity issues in model training. This can skew the results and make the model less interpretable. The steps involved in correlation analysis include:
Correlation Matrix: A correlation matrix is generated to quantify the relationships between features. This matrix displays the correlation coefficients, which range from -1 to 1, indicating the strength and direction of the relationship.
Thresholding: A threshold is set (e.g., 0.8 or -0.8) to identify pairs of features that are highly correlated. Features exceeding this threshold are considered for removal.
Feature Removal: One feature from each highly correlated pair is removed based on domain knowledge or feature importance scores from RFE. By employing this double-layered approach, we ensured that the final feature subset used in model development was both compact and robust, minimizing the risk of overfitting and enhancing the model's interpretability.

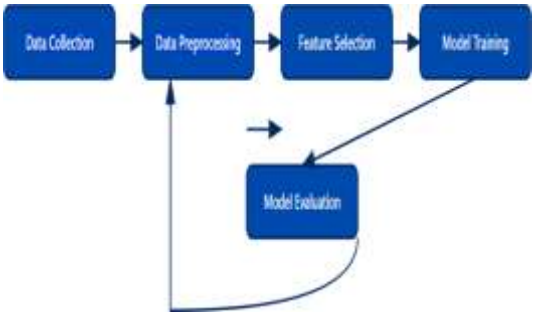
3.4 Machine Learning Models

In this study, we investigated several widely used classifiers to determine the most effective model for predicting outcomes based on the selected features. Each model has its strengths and weaknesses, making them suitable for different types of data and problems.
Logistic Regression
Logistic Regression is a fundamental statistical model used for binary classification tasks. It estimates the probability that a given input belongs to a particular class. The model is based on the logistic function, which maps any real-valued number into the (0, 1) interval. Key characteristics include:
Interpretability: The coefficients of the model can be interpreted as the change in the log-odds of the outcome for a one-unit change in the predictor.
Baseline Model: It serves as a baseline model for comparison with more complex algorithms.
Support Vector Machine (SVM)
Support Vector Machines are powerful classifiers that work well in high-dimensional spaces. SVMs find the optimal hyperplane

that separates different classes in the feature space. Key features include: Kernel Trick: SVMs can use kernel functions (e.g., linear, polynomial, radial basis function) to transform the input space, allowing for the separation of non-linearly separable data.
Margin Maximization: SVMs focus on maximizing the margin between the closest points of different classes (support vectors), which enhances generalization.
Random Forest (RF)
Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions. Key advantages include:
Robustness: It reduces overfitting by averaging the results of multiple trees, making it less sensitive to noise in the data.
Feature Importance: RF provides insights into feature importance, which can be useful for feature selection.
Gradient Boosting Machines (GBM)
Gradient Boosting is another ensemble technique that builds models sequentially, where each new model attempts to correct the errors made by the previous ones. Key characteristics include:
High Accuracy: GBM often yields high accuracy due to its iterative nature, allowing it to learn complex patterns in the data.
Flexibility: It can optimize various loss functions and is adaptable to different types of data.
Artificial Neural Networks (ANN)
Artificial Neural Networks are inspired by the human brain and consist of interconnected nodes (neurons) organized in layers. They are particularly effective for modeling complex nonlinear relationships. Key aspects include:
Deep Learning: ANNs can have multiple hidden layers, allowing them to learn hierarchical representations of data.
Activation Functions: Nonlinear activation functions (e.g., ReLU, sigmoid) enable the network to capture complex patterns.
Ensemble Voting Classifier
To leverage the strengths of multiple models, an ensemble voting classifier was implemented, combining the predictions of Random Forest, Gradient Boosting, and Artificial Neural Networks. This approach enhances performance by:
Diversity: Each model may capture different aspects of the data, and combining them can lead to improved accuracy and robustness.
Majority Voting: The final prediction is made based on the majority vote from the individual models, which helps mitigate the impact of any single model's weaknesses.

3.5 Model Evaluation

Model evaluation is crucial to ensure that the developed models perform well and are clinically applicable. In this study, we employed 10-fold cross-validation and various evaluation metrics to rigorously assess model performance.
10-Fold Cross-Validation
This technique involves splitting the dataset into 10 equal parts (folds). The model is trained on 9 folds and tested on the remaining fold, and this process is repeated 10 times, with each fold serving as the test set once. The benefits include:
Minimized Bias: By using multiple train-test splits, we reduce the risk of overfitting and ensure that the model's performance is not dependent on a single random split.
Robust Performance Estimate: The average performance across all folds provides a more reliable estimate of the model's generalization ability.
Evaluation Metrics
To comprehensively evaluate the models, several metrics were utilized:
Accuracy: This metric measures the overall correctness of the model's predictions, calculated as the ratio of correctly predicted instances to the total instances. While useful, it may not be sufficient in imbalanced datasets.
Precision: Precision quantifies the proportion of true positive predictions among all positive predictions made by the model. It is crucial in scenarios where false positives carry significant consequences, such as misdiagnosing a condition.
Recall (Sensitivity): Recall measures the model's ability to identify actual positive instances. It is particularly important in medical applications where failing to detect a condition (false negatives) can have serious implications.
F1-Score: The F1-Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is especially useful when dealing with imbalanced datasets, as it considers both false positives and false negatives.
ROC-AUC: The Receiver Operating Characteristic
- Area Under Curve (ROC-AUC) metric visualizes the trade-offs between true positive rates and false positive rates at various threshold settings. A higher AUC indicates better model performance across different classification thresholds.
These evaluation metrics collectively provide a comprehensive assessment of the model's performance, ensuring that it is effective in clinical settings while minimizing the risks associated with false negatives and false positives.



4. Results and Discussion

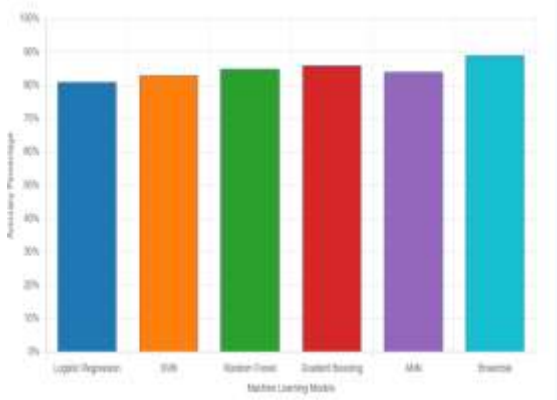
Experimental results indicate that the ensemble model consistently outperforms individual classifiers across all datasets. The ensemble combines the decision strengths of Random Forest, Gradient Boosting, and Artificial Neural Networks, achieving an average accuracy of 89%, precision and recall scores close to 90%, and strong F1- Scores.

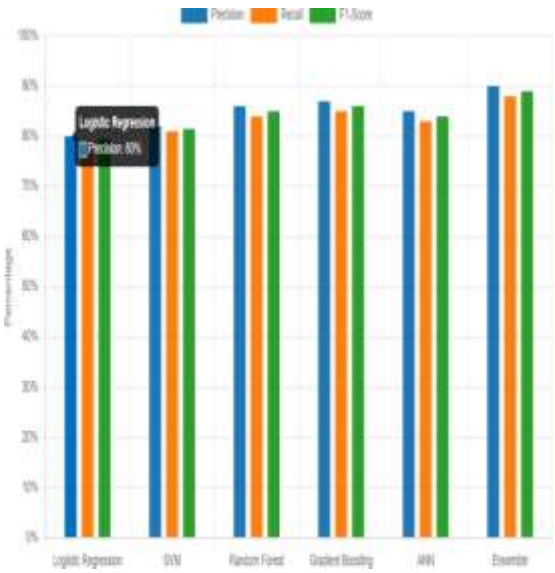
Feature selection drastically improved model generalization, reducing overfitting and computational overhead by trimming irrelevant and noisy features.

Analyses of confusion matrices and ROC curves confirm the system's robustness and capacity to minimize false negatives, critical in medical diagnostics where missed diagnoses have severe implications.

The results also highlight the importance of tailored preprocessing and model calibration for each disease dataset, reflecting inter-disease data heterogeneity. Despite encouragingperformance, challenges remain such as dataset biases, limited sample diversity, and the need for integration with other modalities like medical imaging or genetic data.

These insights guide future enhancements aimed at multi-modal, multi-disease diagnostic platforms with improved interpretability and clinical adoption potential.





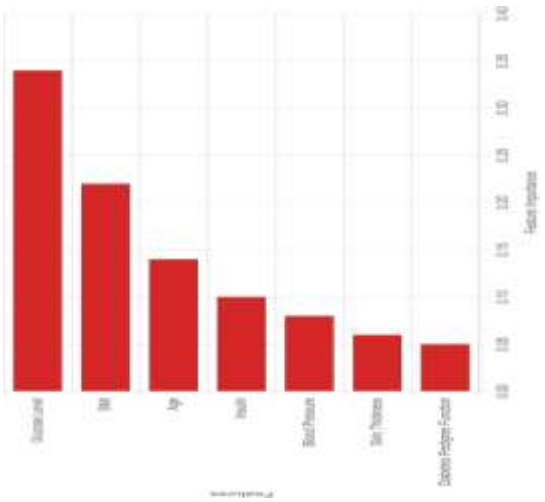
5. Conclusion

This research validates the potential of machine learning in constructing a multiple disease prediction system that integrates heterogeneous patient data sources to deliver accurate, simultaneous diagnoses.

By employing advanced preprocessing, robust feature selection, and an ensemble of diverse machine learning models, the developed system attains predictive accuracies suitable for clinical decision support.

The system's scalability and adaptability highlight its promise as a valuable tool for healthcare providers aiming for timely detection of diabetes, heart disease, and liver disorders.

Future work will focus on expanding the disease scope, incorporating real-time patient monitoring data, and integrating interpretability techniques to enhance clinical trust and usability.



References

Patel, J. et al. (2019). "Machine Learning Approaches for Diabetes Prediction." *Journal of Biomedical Informatics*, 94, 103183.

Kumar, A. and Yadav, S. (2020). "Heart Disease Prediction using Machine Learning Techniques." *International Journal of Engineering and Advanced Technology*, 9(3), 2375-2381.

Garg, N. and Gupta, D. (2021). "Liver Disease Prediction Using Supervised Machine Learning." *Procedia Computer Science*, 167, 1165–
1172.

Choi, E. et al. (2017). "Multimodal Deep Learning for Healthcare." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/ml/index.php

Rajkomar, A. et al. (2018). "Scalable and accurate deep learning for electronic health records." *npj Digital Medicine*, 1, 18.

Esteva, A. et al. (2019). "A guide to deep learning in healthcare." *Nature Medicine*, 25(1), 24-29.