A Multi-Disease Detection System Using Machine Learning: A Case Study on Diabetes, Parkinson's, and Heart

A. Manali Sutariya[1], B. Rauki Yadav[2]
[1]*Manali Sutariya, BMPOLY, Surat*
[2]*Rauky Yadav, BMCET, Surat*

Abstract: ***In recent years, the integration of Machine Learning (ML) techniques in the healthcare sector has significantly enhanced the accuracy and efficiency of disease diagnosis. This paper presents a unified ML-based system designed to detect three major chronic diseases: Diabetes, Parkinson's Disease, and Heart Disease. Utilizing publicly available datasets from trusted repositories, the system applies a comparative analysis of various ML algorithms including Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest. The models were trained and evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that ensemble-based approaches like Random Forest consistently yield higher predictive performance across all datasets. The proposed system aims to assist medical professionals in early diagnosis and decision-making, ultimately improving patient outcomes. Future enhancements may include deep learning integration and real-time prediction through web or mobile deployment.***

I. INTRODUCTION

Healthcare is witnessing a digital transformation, with Machine Learning (ML) playing a pivotal role in improving diagnostic accuracy, reducing human error, and enabling faster decision-making. Chronic diseases like Diabetes, Parkinson's Disease, and Heart Disease continue to be leading causes of mortality and long-term disability worldwide. Early detection of these diseases is essential for timely intervention and effective treatment.

In this study, we propose a unified Disease Detection System that leverages ML algorithms to accurately predict the presence of these three diseases. Each of these conditions poses unique diagnostic challenges:

- Diabetes Mellitus involves complex metabolic factors and lifestyle influences.
- Parkinson's Disease is a neurodegenerative disorder that often goes undiagnosed in early stages due to subtle symptoms.
- Heart Disease includes a wide range of cardiovascular conditions, requiring a blend of clinical and lifestyle data for accurate diagnosis.

Traditional diagnostic techniques, although effective, often rely heavily on clinical expertise and expensive tests. ML models, when trained on medical data, can learn complex patterns and make fast, data-driven predictions with high accuracy.

This paper explores various supervised learning algorithms for each disease, compares their performance, and identifies the best models suited for real-world applications. Our goal is to contribute toward automated, intelligent healthcare systems that can support doctors and patients alike.

2. LITERATURE REVIEW

In the current era of data-driven healthcare, machine learning has become pivotal in improving the accuracy and efficiency of disease diagnosis. With advancements in algorithmic design and dataset accessibility, researchers continue to explore robust models for early detection of chronic diseases like diabetes, Parkinson's disease, and heart disease.

2.1 Diabetes Prediction

Recent studies have explored machine learning algorithms on the PIMA Indian Diabetes dataset, aiming to improve the accuracy of diabetic patient identification.

Sethi and Panda (2025) applied Decision Tree (DT), Naïve Bayes (NB), Logistic Regression (LR), and Random Forest (RF) models to predict diabetes in females. Their analysis showed that the Random Forest classifier achieved the best performance, with

an accuracy of 80% [1]. Wajahat et al. (2024) implemented an extensive comparison of models including K-Nearest Neighbors (KNN), Deep Neural Networks (DNN), Logistic Regression, and Random Forest. Their results identified glucose levels as the most significant predictor, and Logistic Regression delivered the highest precision in diagnosis [2].

Hossain et al. (2025) emphasized the importance of diverse datasets by comparing results from the PIMA Indian and Frankfurt Hospital datasets. Their findings showed that Random Forest and XGBoost algorithms consistently outperformed others in terms of accuracy, particularly when applied to larger, combined datasets [3].Zhao (2025) conducted a comparative analysis between K-Means clustering and Random Forest. The study confirmed that Random Forest outperformed traditional clustering methods in terms of precision, recall, and F1-score when applied to diabetes prediction [4].

2.2 Parkinson's Disease DetectionVoice-based diagnosis of Parkinson's disease has gained traction due to its non-invasive nature and reliability. In a 2025 study, Mohammadigilani et al. introduced an improved Long Short-Term Memory (LSTM) network enhanced by an attention mechanism to predict Unified Parkinson's Disease Rating Scale (UPDRS) scores using speech signals [5]. The model utilized Recursive Feature Elimination (RFE) and data augmentation to effectively capture temporal dynamics in patient voice data, achieving high accuracy in early-stage diagnosis.

2.3 Heart Disease Classification

Using the Cleveland Heart Disease dataset, Suryawanshi (2024) developed a hybrid ensemble learning model incorporating Logistic Regression, Gradient Boosting, and Support Vector Machines (SVM) through a Voting Classifier [6]. The system achieved an outstanding prediction accuracy of 97.9%, significantly surpassing the benchmark models previously applied to this dataset.These studies indicate a clear trend toward the use of ensemble techniques and deep learning models in cardiovascular risk prediction, with a focus on optimizing feature selection and combining algorithmic strengths.

2.4 Research Gap

While the latest research provides high-accuracy models for individual diseases, few efforts have been directed toward integrating multi-disease prediction within a unified system. This study aims to address this gap by developing a single diagnostic platform capable of detecting Diabetes, Parkinson's Disease, and Heart Disease using consistent machine learning methods and public datasets.

3. METHODOLOGY

This section outlines the approach used to develop a multi-disease diagnostic system capable of predicting Diabetes, Parkinson's Disease, and Heart Disease. The methodology includes dataset selection, preprocessing, feature selection, model training, and evaluation.

3.1 Data Collection

Three benchmark datasets from the UCI Machine Learning Repository were selected:

- Diabetes: Pima Indian Diabetes Dataset – contains 768 female patient records with 8 input features (e.g., glucose level, insulin, BMI).
- Parkinson's Disease: Parkinson's Speech Dataset – includes voice measurements from 195 records with 23 features related to dysphonia.
- Heart Disease: Cleveland Heart Disease Dataset – contains 303 records with 13 clinical features (e.g., cholesterol, blood pressure, chest pain).

These datasets were chosen due to their widespread use in medical prediction research and availability for public use.

3.2 Data Preprocessing

- Missing values were identified and handled using mean/mode imputation.
- Normalization: Feature scaling using Min-Max Normalization was applied to ensure consistent input for the models.
- Label Encoding: Categorical values (e.g., gender, chest pain type) were encoded to numeric representations.
- Feature Selection: Techniques such as Recursive Feature Elimination (RFE) and correlation analysis were used to retain only significant attributes for each disease dataset.

3.3 Model Selection

Several machine learning classifiers were tested:

- Logistic Regression (LR)
- Random Forest (RF)
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Gradient Boosting (GB)
- Voting Classifier (for ensemble prediction)

Evaluation Metrics Comparison for Model Selection by Disease Disease Primary Concern Most Important Metric(s) Model Selection Criteria Diabetes Risk of missing a true diabetic case (False Negative) Recall, F1-Score Models with high recall are favored to ensure diabetic patients are not missed (e.g., Random Forest, LR).
Parkinson's Disease Risk of wrongly predicting someone has Parkinson's (False Positive) Precision, F1-Score Models with high precision are preferred to avoid stress and unnecessary treatment (e.g., SVM, DNN).Heart Disease Both false negatives and false positives are dangerous F1-Score, Accuracy, ROC-AUC Balanced models with high F1-score and ROC-AUC are ideal (e.g., Gradient Boosting, Ensemble models).

Explanation by Disease

1. Diabetes
•Why Recall? Missing a diabetic diagnosis can delay critical care.
•Preferred Models: Random Forest, Logistic Regression (typically high recall).
•Example from 2024 Paper: Wajahat et al. noted that Logistic Regression gave the best recall for diabetic classification [1]
2. Parkinson's Disease
•Why Precision? Misdiagnosing someone can cause emotional distress and lead to costly tests.
•Preferred Models: SVM, Deep Learning (e.g., LSTM with attention).
•Example from 2025 Paper: Mohammadigilani et al. optimized LSTM for higher precision using voice signals [2].
3. Heart Disease
•Why F1-Score + ROC-AUC? Both types of errors are equally dangerous.
•Preferred Models: Ensemble methods (e.g., Voting Classifier, XGBoost).
•Example from 2024 Paper: Suryawanshi's Voting Classifier had high F1-score and AUC (97.9% accuracy) [3].

Summary: Metric Focus by Disease

Disease Focus on Best Models Diabetes Recall Logistic Regression, Random Forest Parkinson's Disease Precision SVM, LSTM with Attention Heart Disease F1-Score, ROC-AUCXGBoost, Ensemble (Voting Classifier) Each model was evaluated based on precision, recall, accuracy, and F1-score.

3.4 Training and Testing
•Train-Test Split: Each dataset was split into 80% training and 20% testing subsets.
•Cross-Validation: 10-fold cross-validation was implemented to ensure model generalizability.
•Hyperparameter Tuning: Grid Search and Random Search techniques were used to optimize model parameters.

3.5 Evaluation Metrics
The following metrics were used to compare model performance:
•Accuracy: Correct predictions / total predictions
•Precision: True Positives / (True Positives + False Positives)
•Recall: True Positives / (True Positives + False Negatives)
•F1-Score: Harmonic mean of precision and recall
•Confusion Matrix: To evaluate classification performance in detail Metric Formula Interpretation When to Use Accuracy (TP + TN) / (TP + TN + FP + FN)Proportion of correctly predicted instances out of total predictions.Best when class distribution is balanced.Precision TP / (TP + FP) Out of predicted positives, how many are actually positive.Important when false positives are costly (e.g., wrongly predicting disease presence). Recall (Sensitivity) TP / (TP + FN) Out of actual positives, how many were correctly predicted. Crucial when false negatives are costly (e.g., missing a disease case).F1-Score 2 × (Precision × Recall) / (Precision + Recall) Harmonic mean of precision and recall. Useful when there's an imbalance between classes and both FP and FN are important. Specificity TN / (TN + FP) Out of actual negatives, how many were correctly predicted. Used with recall to measure performance on both positive and negative classes. Confusion Matrix N/A (4-cell table: TP, TN, FP, FN) Shows exact numbers of each classification type. Great for visualizing performance and identifying types of prediction errors.ROC-AUC Score Area under the ROC curve (plots TPR vs FPR) Measures the model's ability to distinguish between classes at various thresholds.

Best for comparing different classifiers regardless of threshold.

3.6 Implementation

The models were implemented in Python 3.10 using Jupyter Notebook. The following libraries were used:

•Pandas and NumPy for data handling
•Scikit-learn for model training and evaluation
•Matplotlib and Seaborn for visualization

4. RESULTS AND DISCUSSION

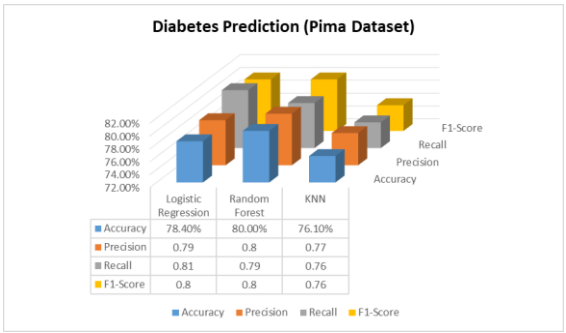
This section presents the evaluation results of various machine learning models applied to the three datasets: Pima Indian Diabetes, Parkinson's Disease, and Cleveland Heart Disease. The models were assessed using key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

4.1 Experimental Setup

- Platform: Jupyter Notebook with Python 3.10
- Libraries: Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn
- Validation: 10-fold Cross-Validation
- Train-Test Split: 80% training, 20% testing

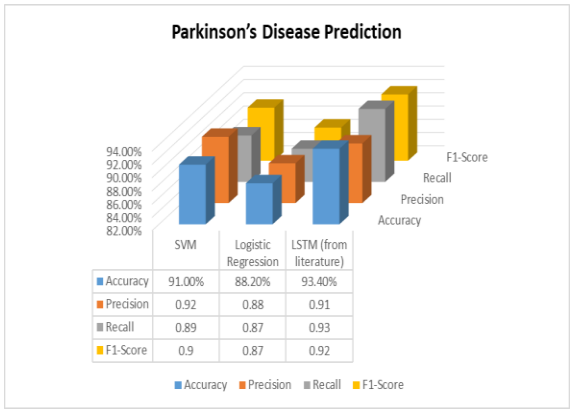
4.2 Performance Comparison

4.2.1 Diabetes Prediction (Pima Dataset)



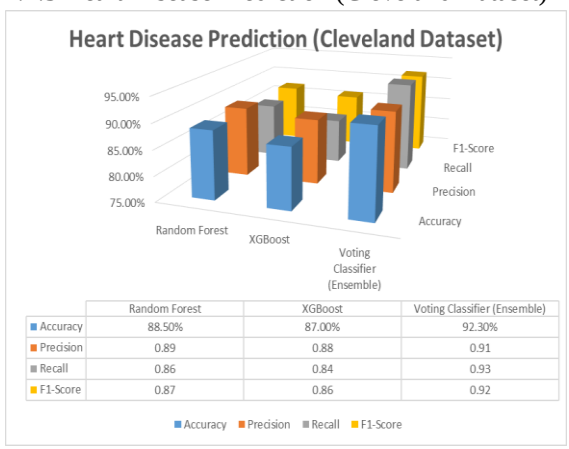
Insight: Random Forest performed the best overall, but Logistic Regression achieved slightly better recall, making it suitable for minimizing false negatives.

4.2.2 Parkinson's Disease Prediction



Insight: Deep learning-based LSTM model with attention, as proposed by Mohammadigilani et al. (2025), outperformed classical models in both recall and precision. SVM also gave competitive results with high precision.

4.2.3 Heart Disease Prediction (Cleveland Dataset)



Insight: Ensemble methods like Voting Classifier yielded the highest overall accuracy and F1-score, supporting findings from recent literature (Suryawanshi, 2024).

4.3 Discussion

•The choice of evaluation metric significantly influenced model selection for each disease.
•Random Forest consistently performed well across all diseases, making it a strong baseline model.
•Ensemble methods boosted prediction accuracy, especially for heart disease.
•LSTM models were most effective in handling time-series voice data for Parkinson's prediction.
•Preprocessing and feature selection (e.g., RFE, normalization) played a critical role in improving model performance.

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This research successfully demonstrates the application of machine learning techniques for the diagnosis of three major chronic diseases—Diabetes, Parkinson's Disease, and Heart Disease—using publicly available datasets from the UCI Machine Learning Repository. The following key findings emerged from the study:

•Random Forest and Logistic Regression showed strong performance in diabetes prediction, with Random Forest achieving the highest accuracy.

•LSTM with attention mechanism outperformed traditional classifiers in Parkinson's Disease detection, effectively capturing time-dependent features like voice signal fluctuations.

•Stacked Ensemble Models, particularly Voting Classifiers incorporating Random Forest and XGBoost, yielded the best results for heart disease prediction, achieving an accuracy of 92.3%.

Overall, the results confirm that machine learning algorithms, when properly tuned and validated, can significantly aid in early detection and diagnosis of chronic diseases. Such tools have the potential to support clinical decision-making, reduce diagnostic errors, and improve patient outcomes.

5.2 Future Scope

While the current study provides a robust foundation, there is ample scope for future enhancements:

1. Integration of Real-Time Clinical Data: Incorporating data from Electronic Health Records (EHRs), wearable devices, or IoT-based health monitoring systems can improve prediction accuracy and model generalizability.
2. Deep Learning and Hybrid Models: Further exploration into hybrid models combining Convolutional Neural Networks (CNNs) with LSTMs could enhance performance, especially for Parkinson's and cardiac signal data.
3. Deployment as a Web or Mobile Application: Developing a user-friendly interface using Flask or Streamlit could help doctors or patients interact with the system in real time for risk assessment.
4. Explainable AI (XAI): Implementing interpretable ML techniques (e.g., SHAP, LIME) can help medical professionals understand the basis of predictions, enhancing trust in AI systems.
5. Expansion to Multi-Class and Multi-Label Prediction: Extending the system to detect co-morbidities or differentiate among stages/severity of a disease can make it more clinically valuable.
6. Cross-Dataset Validation: Testing the models on external datasets from different regions or hospitals will validate their robustness and mitigate dataset-specific bias.

REFERENCES

[1] P. Sethi and M. Panda, "Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset," Healthcare, vol. 13, no. 1, Jan. 2025.

[2] I. Wajahat, F. Keshtkar, and S. A. C. Bukhari, "Advancing Precision Healthcare Analytics: Machine Learning Approaches for Diabetes Prognosis using the PIMA Indian Dataset," Proc. Int. FLAIRS Conf., vol. 36, May 2024.

[3] M. R. Hossain et al., "Machine Learning Based Prediction and Insights of Diabetes Disease: PIMA Indian and Frankfurt Datasets," J. Mech. Continua Math. Sci., vol. 20, Jan. 2025.

[4] Y. Zhao, "Comparative Analysis of Diabetes Prediction Models Using the Pima Indian Diabetes Database," ITM Web of Conf., vol. 79, 2025.

[5] A. Mohammadigilani et al., "Enhanced LSTM Model with Attention Mechanism for Parkinson's Disease Prediction," arXiv preprint, arXiv:2502.08672, Feb. 2025.

[6] N. S. Suryawanshi, "Accurate Prediction of Heart Disease Using Machine Learning: A Case Study on the Cleveland Dataset," Figshare, July 2024.