



## Disease Identification Virtual Yantra (DIVY)

*P. Swati<sup>1</sup>, Pranav Sahu<sup>2</sup>, Tushar Pandey<sup>3</sup>, Arya Sharma<sup>4</sup>, Kunal Sahu<sup>5</sup>*

<sup>1</sup>Assistant Professor, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

<sup>2,3,4,5</sup> Student, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

### ABSTRACT

Disease Identification Virtual Yantra (DIVY) is a machine learning-based framework for predicting multiple diseases—diabetes, heart disease, and liver disease. Using publicly available datasets, preprocessing, feature engineering, and machine learning models such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting were applied. Evaluation metrics included accuracy, precision, recall, F1-score. Results showed that ensemble models and neural networks achieved high diagnostic accuracy. The findings highlight the potential of integrated machine learning frameworks in supporting healthcare diagnostics. Future work will focus on improving model interpretability and expanding dataset diversity.

Keywords: Machine Learning, Disease Prediction.

### I. INTRODUCTION

Chronic diseases such as diabetes, heart disease, and liver disease are major contributors to global health burdens. Early diagnosis is critical for effective treatment and better patient outcomes. Traditional diagnostic methods are often labour-intensive and prone to delays. Machine learning (ML) offers a promising solution for scalable, accurate, and efficient disease prediction. Disease Identification Virtual Yantra (DIVY) helps in early diagnosis of these diseases.

#### 1. Background

Diabetes, heart disease, and liver disease are among the most prevalent non-communicable diseases worldwide. According to the World Health Organization, diabetes affects over 422 million people, while cardiovascular diseases account for 17.9 million deaths annually. Liver diseases, including cirrhosis and hepatitis, are significant contributors to global morbidity and mortality.

#### 2. Motivation

Although machine learning (ML) has been widely applied to individual diseases, integrating predictions for multiple diseases within a unified framework remains underexplored. A multi-disease system could streamline diagnostics, particularly in regions with limited healthcare resources.

### II. RELATED WORK

#### 1. Diabetes Prediction

Numerous studies utilize Diabetes Dataset to develop models such as Logistic Regression and Random Forest for diabetes prediction. Ensemble methods often outperform simpler models, achieving accuracies above 97%.

#### 2. Heart Disease Prediction

Heart disease prediction has been widely studied using datasets like the Cleveland Heart Disease Dataset. Ensemble learning models, such as Gradient Boosting, have achieved excellent precision and recall in predicting cardiovascular risks.

#### 3. Liver Disease Prediction

Liver disease, including conditions like cirrhosis and hepatitis, is increasingly diagnosed using ML models. The dataset has been extensively used in research. Random Forest and Gradient Boosting Classifier are commonly employed, achieving high accuracy.

### III. METHIODOLOGY

#### 1. Datasets

Kaggle Diabetes Dataset: Features include glucose levels, BMI, and insulin levels.

Kaggle's Heart Disease Dataset: Includes cholesterol levels, ECG results, and maximum heart rate.

Kaggle's Liver disease Dataset: Features such as bilirubin, alkaline phosphatase, albumin, and aspartate transaminase.

#### 2. Preprocessing

Missing values were imputed using mean substitution.

Continuous features were normalized using Min-Max scaling.

#### 3. Feature Engineering

Recursive Feature Elimination (RFE) and mutual information were used for feature selection.

Domain-specific preprocessing was applied to liver disease datasets, including log transformations for enzyme levels.

#### 4. Algorithms

**Baseline Models:** Logistic Regression for performance benchmarking.

**Advanced Models:** Random Forest, SVM with RBF kernel, Gradient Boosting, and Multi-layer Perceptron (MLP).

**Hyperparameter Tuning:** Grid search was performed for parameters like tree depth (Random Forest) and learning rate (Gradient Boosting such as LightGBM which is a gradient boosting framework optimized for speed and accuracy).

#### 5. Evaluation Metrics

The models were evaluated using:

- (a) Accuracy.
- (b) Precision, Recall, and F1-Score.

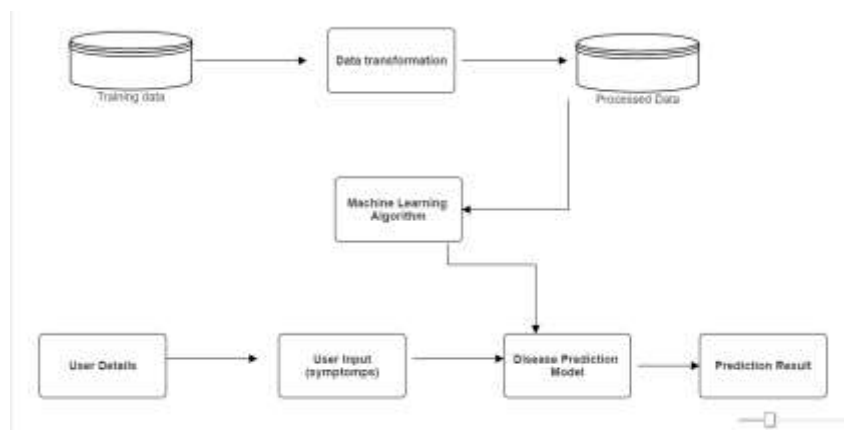


Fig.1 Flow Chart

### VI. RESULT

#### 1. Diabetes Prediction

Best Model: Random Forest achieved an accuracy of 97%.

Key Insight: Glucose and BMI were the most important predictive features.

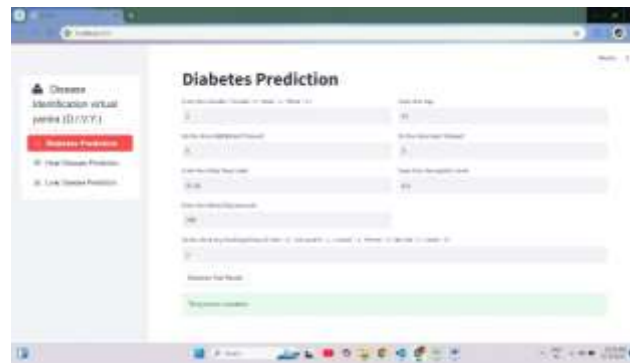


Fig.2 Diabetes output

## 2. Heart Disease Prediction

Best Model: Gradient Boosting achieved an accuracy of 91.2%.

Key Insight: Cholesterol and resting ECG results were key contributors to predictions.

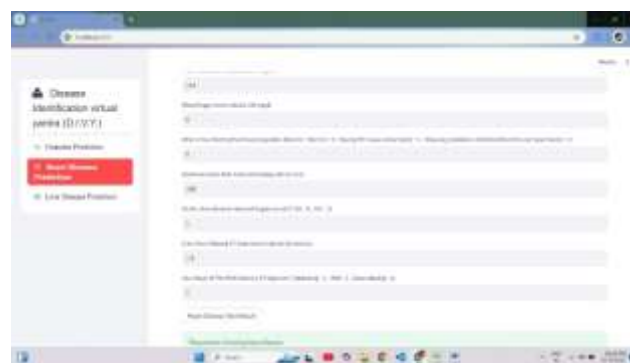


Fig.3 Heart Disease output

## 3. Liver Disease Prediction

Best Model: Gradient Boosting achieved an accuracy of 90%.

Key Insight: BMI, Liver function test, Hypertension were the most critical features, highlighting their diagnostic significance.

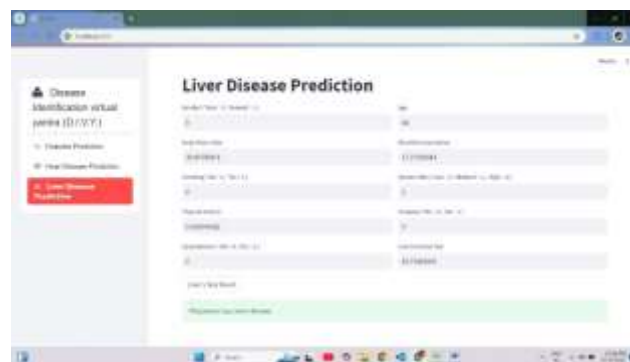


Fig.4 liver Disease output

## 4. Comparative Analysis

Ensemble models consistently performed well across diseases.

## 5. Error Analysis

**Diabetes:** False negatives occurred in borderline glucose cases, suggesting additional lifestyle features could improve predictions.

**Heart Disease:** Atypical symptoms led to false positives, indicating potential dataset bias.

**Liver Disease:** Misclassification occurred for patients with overlapping conditions such as hepatitis and jaundice.

---

## V. CONCLUSION

This study demonstrates the effectiveness of machine learning models in predicting diabetes, heart disease, and liver disease. The findings underline the potential of ML to enhance diagnostic accuracy and efficiency. Ensemble methods and neural networks emerged as the most reliable approaches across all datasets. Future work will focus on testing the models in clinical environments and incorporating additional features such as genetic and lifestyle data.

### Future Directions

- (a) **Explainable AI:** Incorporating SHAP or LIME to improve model interpretability.
- (b) **Real-World Testing:** Validating models on clinical datasets to assess generalizability.
- (c) **Integration with Wearables:** Leveraging continuous data from wearable devices for real-time disease monitoring.
- (d) **Personalized Medicine:** Adding genetic and demographic data to enhance prediction accuracy.

### References

---

- [1] Kumar Bibhuti B. Singh, Ashutosh Sharma , Ashish Verma ,Ranjeet Maurya, Dr. Yusuf Perwej,"Machine Learning for the Multiple Disease Prediction System," IISRCSELT,2024.
- [2] Parshant , Dr. Anu Rathee," Multiple Disease Prediction Using Machine Learning," IRE Journals,2023.
- [3] Samrat Kumar Dey, Ashraf Hossain, Md. Mahbubur Rahman, "Implementation of a Web Application to Predict Diabetes Disease:An Approach Using Machine Learning Algorithm", *2018 21st International Conference of Computer and Information Technology (ICCIT)*.
- [4] Vansika Gupta et al., "Multiple Disease Prediction Using Machine Learning," International Journal for Multidisciplinary Research (IJFMR), 2024
- [5] IEEE, "Multiple Disease Prediction System Using Machine Learning," IEEE Xplore, 2024.
- [6] IEEE, "Multiple Disease Prediction Using Machine Learning and Deep Learning with Web Technology," IEEE Xplore, 2024