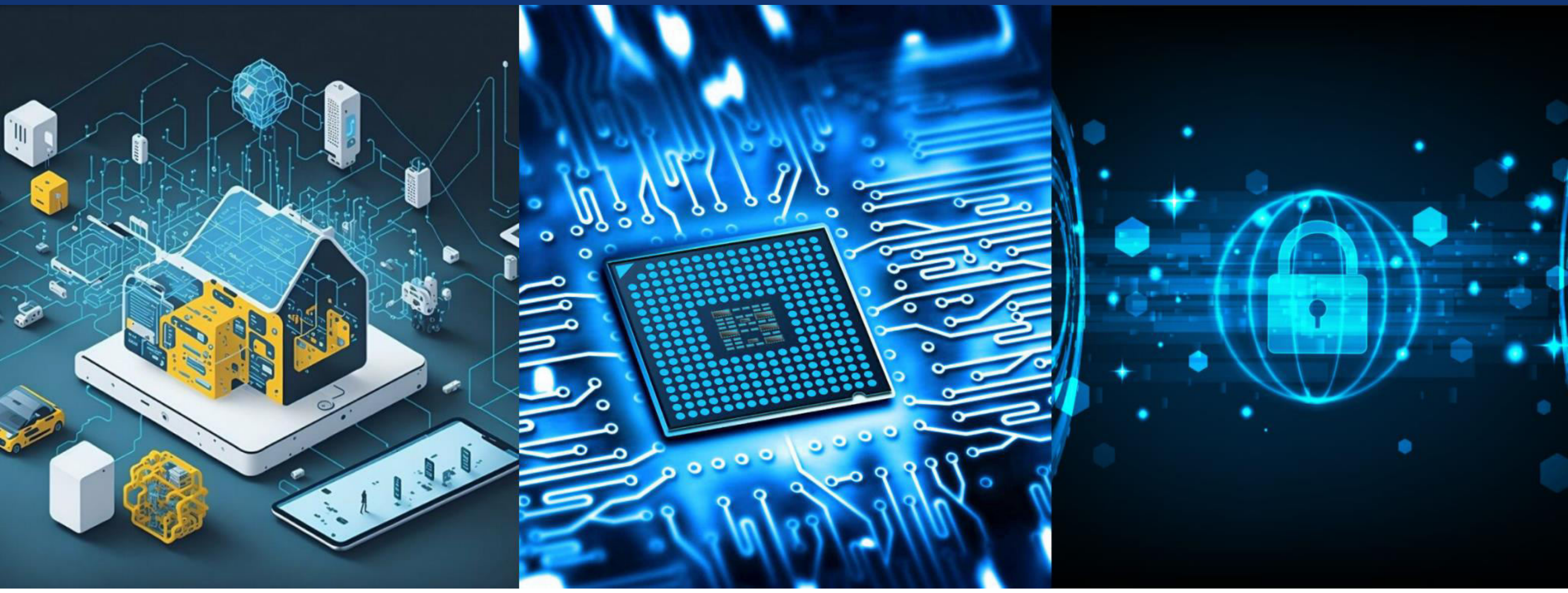
IJIRCCE

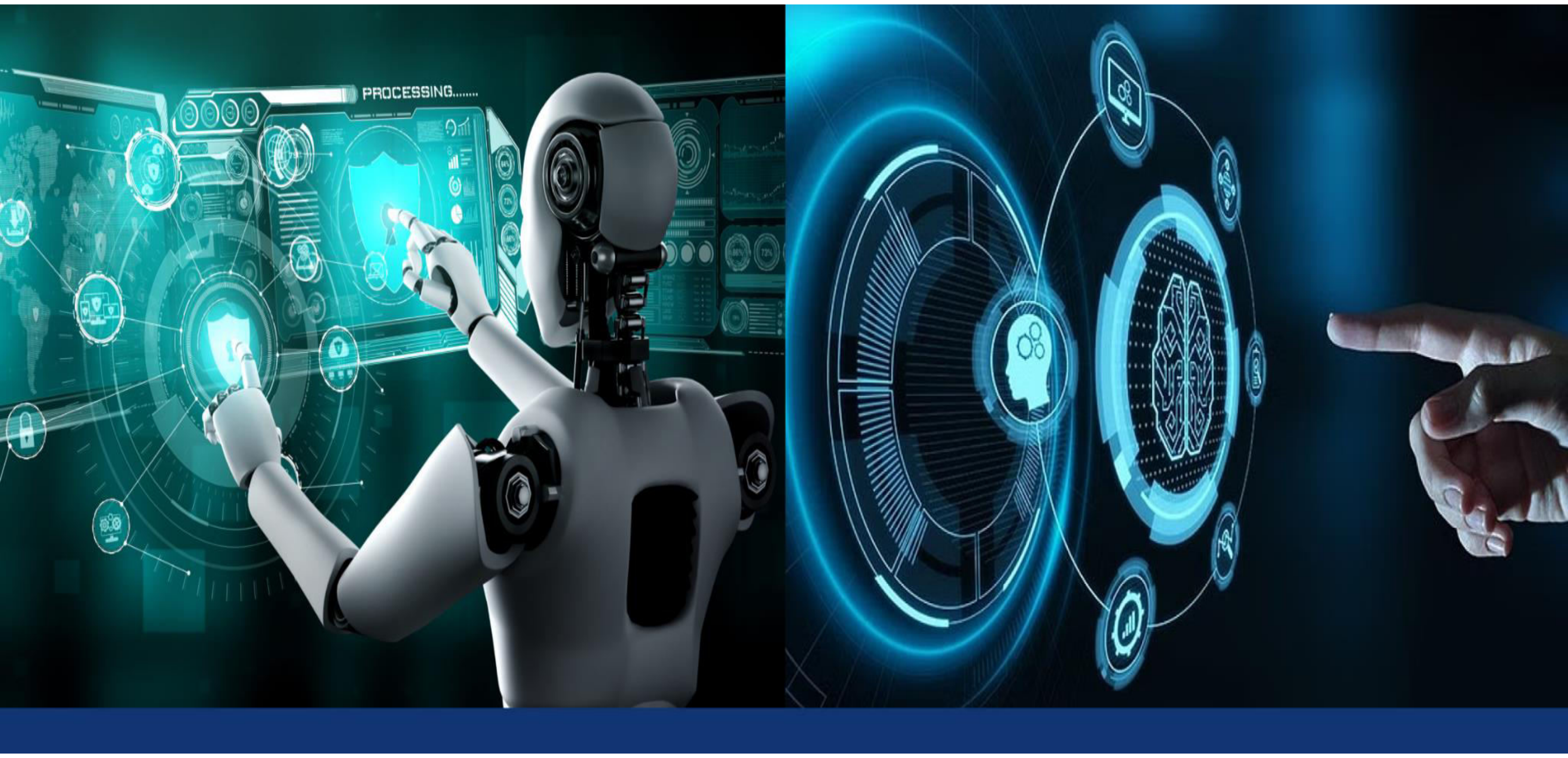
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 3, March 2025

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Liver, Kidney, Heart and Diabetes Disease Prediction using Machine Learning

Sanket Pehere[1], Shubham Gadekar[2], Utkarsh Kulkarni[3], Ramdas Bagawde[4]

Government College of Engineering and Research, Avasari Khurd, Pune, India[1,2,3,4]

ABSTRACT: Chronic illnesses like liver disease, kidney disease, heart disease, and diabetes present major health difficulties across the globe, making timely identification critical for effective management and prevention. Traditional diagnostic methods rely on clinical evaluations, which can be time-consuming, costly, and prone to human error. To address this challenge, we propose a predictive system for liver, kidney, heart, and diabetes conditions utilizing machine learning, employing models such as Logistic Regression, Random Forest, SVM, KNN, AdaBoost, Gradient Boosting, and Decision Trees. This system analyzes key health metrics, including blood pressure, glucose levels, cholesterol, liver function tests, and lifestyle factors to determine disease risk. A structured data preprocessing pipeline guarantees high-quality input by handling missing values, encoding categorical variables, and detecting outliers. The models are evaluated using metrics like accuracy, precision, recall, and F1-score to ensure their reliability. Developed as a web application with Flask, this system provides a user-friendly interface for real-time predictions, coupled with a SQL database for the storage of patient information. This AI-driven solution seeks to help healthcare professionals and individuals make informed health decisions, improving early detection and preventive strategies.

KEYWORDS: Modelling, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours Algorithm (KNN), Deployment, Exploratory Data Analysis, Hypothesis Development, Feature Engineering, Preprocessing Data, and Supervised Learning.

I. INTRODUCTION

The prevalence of chronic diseases such as Heart Disease, Kidney Disease, Liver Disease, and Diabetes is rising at an alarming rate, posing a significant public health challenge. In India alone, the annual deaths from heart disease increased from 2.26 million in 1990 to 4.77 million in 2020 [1]. Kidney disease has emerged as a critical concern due to its high prevalence and mortality rates [2]. Liver disease accounted for 18.3% of global liver-related deaths in 2015, highlighting its severe impact [3]. Additionally, diabetes cases in India surged to 77 million in 2019 and are expected to exceed 134 million by 2045, with 57% of cases remaining undiagnosed [4]. As the number of patients continues to grow, the burden on the healthcare system increases, making early and accurate disease prediction essential for effective intervention.

This system utilizes machine learning (ML) techniques to predict the likelihood of these diseases based on patient health records. By leveraging large datasets and advanced classification models, the system ensures efficient processing and high accuracy in disease prediction. Machine learning enhances early diagnosis, personalized risk assessment, and decision-making, reducing the burden on healthcare professionals. Compared to traditional diagnostic methods, ML-driven approaches are faster, more precise, and scalable, making them invaluable in preventive healthcare. By integrating AI-driven analytics, this project aims to develop an intelligent and clinically relevant diagnostic system, ultimately improving healthcare accessibility and patient outcomes.

This system leverages a dataset containing patient health records and applies various classification models to assess risk factors effectively.

A. Data Collection:

We have collected the dataset from Kaggle platform. Kaggle is a publicly available data repositories and research databases [5]. Also some data is also collected from nearby Hospitals. This dataset consists of various clinical attributes such as:

Demographic Information: Age, Gender

Vital Signs: Blood Pressure (Systolic & Diastolic)
Biochemical Indicators: Cholesterol, Glucose Levels
Lifestyle Factors: Smoking, Alcohol Consumption, Physical Activity

B. Data Pre-processing [6]:
To ensure data quality and enhance model performance, the following preprocessing steps were applied:
a) Handling Missing Values: Median Imputation: Median Imputation is a technique in which the missing values in a variable with the median value of that variable. It is used for continuous numerical values (e.g., BP, cholesterol).
b) Label Encoding : In this technique unique number is assigned to each category in a dataset. Here it is used for binary categorical variables (e.g., Smoking: Yes/No → 1/0).
c) One-Hot Encoding : It converts categorical values into binary vectors (dummy variables).
d) Outlier Detection & Removal: Z-score Method: It is a statistical technique which measures how far a data point is from the mean in terms of standard deviations. Here outliers were identified and capped based on the standard deviation threshold.

C. Feature Extraction [7]:
Feature extraction plays a critical role in improving model efficiency by deriving informative features from raw data. Statistical Feature Extraction is a type of feature extraction where descriptive statistics is computed from raw data to capture essential information and patterns. It helps ML models to better understand the dataset. In this project Mean, Median, and Standard Deviation Calculations are used to capture central tendencies and variations in blood pressure, cholesterol, and glucose levels.

D. Model Training [8]: To develop a robust predictive system, multiple supervised learning classifiers were trained and evaluated.
The selected models include:
a) Logistic Regression (LR): It is a supervised Machine Learning algorithm used for binary classification problems. Used as a baseline model to understand feature importance.
b) Random Forest (RF):It can be used for both classification and regression. It combines multiple decision trees to improve accuracy and reduce overfitting. An ensemble-based decision tree model that enhances predictive performance.
c) Support Vector Machine (SVM):This ML Algorithm also can be used for both classification and regression. This algorithm is highly effective for binary classification and works well when data is linearly when data is linearly separable. Maps input data into higher-dimensional space to find the best separation boundary.
d) K-Nearest Neighbors (KNN):It is a non parametric, instance based algorithm used for classification and regression. A distance-based classifier that predicts outcomes based on the majority vote of neighboring data points.
e) AdaBoost Classifier: This is an ensemble learning algorithm that merges several weak learners to form a strong classifier. f) Gradient Boosting Classifier: This is an ensemble learning method that constructs a robust classifier by reducing the error through gradient descent.
g) Decision Tree (DT): It splits data into branches based on feature values, to form a tree like structure. This tree-structured algorithm that classifies data by making sequential decisions based on feature values.

Each model was evaluated using standard performance metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC to determine its effectiveness.

E. Model Testing & Deployment:
a) Model Training and Hyperparameter Tuning
Each model was trained using the training dataset (80%), with hyperparameter tuning applied to optimize performance. The following techniques were used:
Cross-Validation: Prevented overfitting by evaluating models on multiple train-test splits.

b) Model Testing and Evaluation
Models were tested using the 20% test dataset, and their performance was measured using the following metrics:

- Accuracy: Measures overall correctness of predictions.
- Precision: Evaluates how many predicted positive cases were actually positive.

- Recall (Sensitivity): Measures how well the model identifies true positive cases.
- F1 Score: Balances Precision and Recall.
- AUC-ROC Score: Determines the model’s ability to distinguish between classes.

c) Model Deployment

To make the trained models accessible, the system was deployed using a Flask-based web application.

The deployment process included:

Model Serialization: The trained models were saved as .pkl files using joblib for efficient storage and retrieval.

Backend Implementation: A Flask API was developed with endpoints to accept user input, process data, and return predictions in real-time.

Frontend Development: A simple HTML, CSS, JavaScript, and Bootstrap-based frontend was designed for user interaction.

Integration with SQL: User inputs and predictions were logged in a database for further analysis.

II. RELATED WORK

Paper Title	Authors (Year)	Algorithm/Strategy Used	Advantages	Disadvantages
Survey of Heart Disease Prediction and Identification using Machine Learning Approaches [9]	Ramya G. Franklin, Dr. B. Muthukumar (2020)	CNN, LSTM, Naïve Bayes, SVM	This study explores multiple machine learning techniques, providing a comprehensive comparison and achieving improved prediction accuracy.	The approach is computationally expensive and requires large datasets for optimal performance.
Heart Disease Prediction using Machine Learning Techniques [2] [10]	Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta (2020)	Random Forest, SVM, Naïve Bayes, Decision Tree	The study highlights that the Random Forest model achieves the highest accuracy, making it a reliable and interpretable choice for heart disease prediction.	The model’s performance is highly dependent on the quality and diversity of the dataset used for training.
Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyperparameter Tuning [3] [11]	Sateesh Ambesange, Vijayalaxmi A, Sridevi S, Dr. Venkateswaran, Dr. Yashoda B S (2020)	Logistic Regression with feature selection, PCA, Grid Search	The model achieves high accuracy by optimizing feature selection, reducing overfitting, and improving predictive performance.	The approach requires careful hyperparameter tuning, which can be complex and time-consuming.
Diabetes Prediction Using Machine Learning [4] [12]	KM Jyoti Rani (2020)	KNN, Logistic Regression, Random Forest, SVM, Decision Tree	The study provides a comparison of multiple models to determine the optimal classifier for diabetes prediction.	Model accuracy varies significantly depending on the dataset used, affecting consistency across different data sources.
Diabetes Prediction using Machine Learning Algorithms [5] [13]	Aishwarya Mujumdar, Dr. Vaidehi V (2019)	Big Data Analytics with ML models	The approach enhances classification accuracy by integrating external factors into the prediction process.	Requires large-scale data processing capabilities, making implementation computationally

				demanding.
Chronic Kidney Disease Prediction based on Machine Learning Algorithms [6] [14]	Md. Ariful Islam, Md. Ziaul Hasan Majumder, Md. Alomgeer Hussein (2023)	XGBoost and other ML classifiers	The proposed model achieves high accuracy (98.3%) and demonstrates robustness in chronic kidney disease prediction.	The model is computationally intensive and requires significant processing power for training and inference.
Chronic Kidney Disease Prediction Using Machine Learning Methods [7] [15]	Imesh Udara Ekanayake, Damayanthi Herath (2020)	Extra Trees, Random Forest	The study identifies the most accurate models for CKD prediction, ensuring reliability for medical applications.	The model's effectiveness depends on well-preprocessed data, requiring extensive feature engineering.

This Liver, Kidney, Diabetes and Heart Disease Prediction Using ML system addresses key limitations observed in existing research by providing a unified approach to predicting multiple diseases (Heart, Liver, Kidney, and Diabetes), whereas most studies focus on a single condition [16]. By integrating real-world hospital data with publicly available Kaggle datasets, the system ensures improved generalization and accuracy across diverse patient populations. Additionally, advanced data preprocessing techniques, including Median Imputation, Z-score outlier detection, and Feature Encoding, enhance data quality and mitigate issues related to missing values and inconsistencies, which are common challenges in prior studies. Unlike conventional models that rely on individual classifiers such as Logistic Regression or Decision Trees, this system leverages ensemble learning techniques (Random Forest, SVM, KNN, AdaBoost, Gradient Boosting) with hyperparameter tuning, leading to higher predictive performance. Furthermore, the system is deployed as a Flask-based web application, enabling real-time disease prediction, user accessibility, and database integration for logging and analyzing predictions. Future advancements, including real-time data integration from wearable health devices, deep learning implementation, and correlation analysis between diseases with shared symptoms, enhance the scalability and applicability of the system, making it a robust tool for modern healthcare diagnostics.

III. METHODOLOGY

This Multiple Disease Prediction System adheres to a systematic methodology to guarantee precise and effective disease forecasting. It consists of data collection, preprocessing, model training, evaluation, and deployment to establish a sturdy system capable of predicting Heart, Liver, Kidney, and Diabetes diseases.

A. Data Collection: The dataset for this project is obtained from Kaggle and local hospitals, ensuring variety and real-world relevance. It comprises critical patient characteristics such as:

- Demographic Information: Age, Gender.
- Vital Signs: Blood Pressure (Systolic and Diastolic).
- Biochemical Indicators: Cholesterol, Glucose Levels.
- Lifestyle Factors: Smoking, Alcohol Consumption, Physical Activity.

By incorporating real-world hospital data, the system guarantees enhanced accuracy and flexibility across various patient demographics.

B. Data Preprocessing: To improve model performance, the dataset is subjected to thorough preprocessing, which includes:

- Handling Missing Values: Median Imputation is utilized to address missing values in numerical characteristics such as blood pressure and cholesterol.
- Feature Encoding:
    - Label Encoding is implemented for binary categorical variables (e. g. , Smoking: Yes/No → 1/0).

- One-Hot Encoding is applied for multi-category features to generate binary representations.

• Outlier Detection and Removal: The Z-score method is employed to identify and cap extreme values, ensuring consistent model performance.

• Feature Extraction: Statistical measures (Mean, Median, Standard Deviation) are calculated for attributes like blood pressure, cholesterol, and glucose levels to capture significant patterns.

C. Model Training and Evaluation: To guarantee optimal disease prediction, various supervised machine learning models are trained and assessed:

• Logistic Regression: Used as a foundational model for binary classification.
• Random Forest: An ensemble learning model that minimizes overfitting and enhances accuracy.
• Support Vector Machine (SVM): Effective for high-dimensional datasets and determining optimal decision boundaries.
• K-Nearest Neighbors (KNN): A non-parametric model that forecasts outcomes based on adjacent data points.
• AdaBoost Classifier: Amplifies weak classifiers to improve overall predictive performance.
• Gradient Boosting Classifier: Utilizes sequential learning to reduce errors and boost accuracy.
• Decision Tree: A straightforward, interpretable model that categorizes data through hierarchical decision processes.

To enhance performance, Grid Search and Random Search are employed for hyperparameter optimization. The models are assessed using Accuracy, Precision, Recall, F1-Score, and AUC-ROC Score. Moreover, cross-validation is utilized to avert overfitting and enhance generalization.

D. Model Deployment: The highest-performing model is implemented as a Flask-based web application to offer real-time disease predictions. The deployment process involves:

• Backend Implementation: A Flask API handles user input and returns disease predictions.
• Frontend Development: An intuitive interface is developed using HTML, CSS, JavaScript, and Bootstrap.
• Database Integration: An SQL database records user inputs and predictions for ongoing analysis and system enhancements.

This methodology guarantees that the Multiple Disease Prediction System Using ML is precise, scalable, and user-friendly, rendering it a valuable resource in contemporary healthcare.

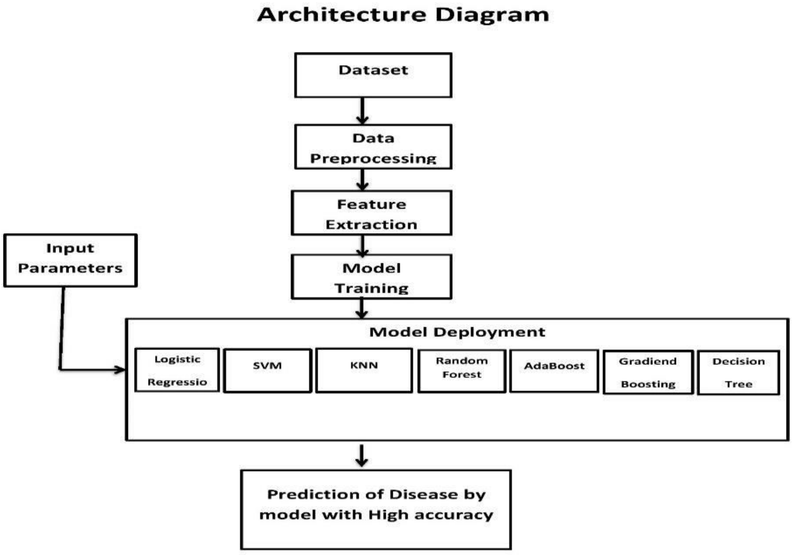


Figure 1: Architechture Diagram

IV. RESULTS AND DISCUSSION

This Multiple Disease Prediction System evaluates the performance of various machine learning models for predicting four diseases: Diabetes, Heart Disease, Kidney Disease, and Liver Disease. The models assessed include Logistic Regression, Random Forest, SVM, Decision Tree, Gradient Boosting, AdaBoost, and KNN. The precision metric is used to compare their effectiveness.

Overall Model Performance

A. Best Performing Models:
   - Random Forest and Gradient Boosting consistently achieve high precision across all diseases, making them the most reliable models for multi-disease prediction.
   - Gradient Boosting and Random Forest achieve 100% precision for Kidney Disease, showing their robustness in handling complex patterns.

B. Moderate Performers:
   - AdaBoost and KNN also perform well, maintaining precision above 90% for most diseases, but show some variations for Heart Disease.

C. Weaker Performers for Certain Diseases:
   - Logistic Regression and SVM struggle with Heart Disease, achieving 75.63% and 66.10% precision, respectively.
   - Decision Tree shows low precision for Diabetes (72.77%), indicating it may not be the best choice for this particular classification.

Significance of Results

- The system effectively uses ensemble learning techniques (Random Forest, Gradient Boosting, AdaBoost) to improve overall prediction accuracy.
- Models like SVM and Decision Tree, which perform poorly for certain diseases, can be improved by feature engineering or combined with other models in an ensemble approach.
- The results highlight that no single model is universally best, but Random Forest and Gradient Boosting are the most stable across multiple diseases.

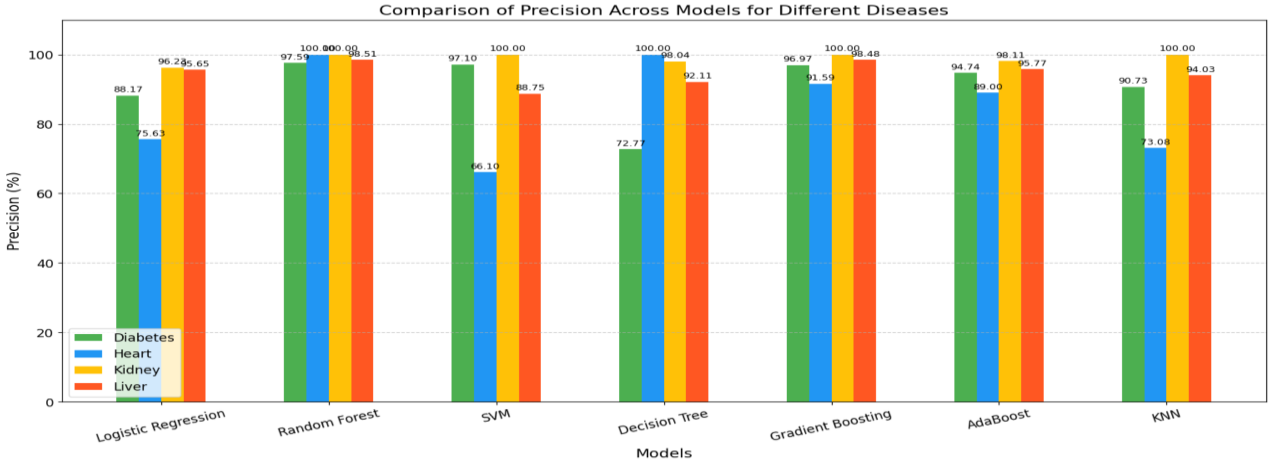


Figure 2 : Comparison of Precision across Models for Different Diseases

V. CONCLUSION

This research demonstrates the effectiveness of machine learning in disease risk prediction, particularly using models such as Random Forest, Decision Tree, and Support Vector Machine (SVM), which have exhibited high predictive accuracy. By systematically incorporating data preprocessing, feature extraction, model training, evaluation, and deployment, a web-based prediction system was developed to assist in early diagnosis. The system enhances preventive healthcare by supporting timely medical intervention and reducing reliance on costly diagnostic procedures. With its implementation as a Flask-based web application, it provides real-time disease predictions with a user-friendly interface and SQL database integration, making it accessible for both patients and healthcare professionals.

VI. FUTURE WORK

This project has significant potential for expansion into a more advanced diagnostic tool capable of predicting a broader range of diseases beyond heart, liver, kidney, and diabetes. Future improvements include:

1. Expanding the dataset – Incorporating larger and more diverse datasets to enhance accuracy and generalization across different populations.
2. Optimizing algorithms – Integrating deep learning models and ensemble techniques to improve predictive performance.
3. Enhancing user experience – Developing a more intuitive and interactive interface for better usability.

With advancements in data integration, model optimization, and AI-driven insights, this system has the potential to revolutionize healthcare by enabling early disease detection, reducing diagnostic workload, and improving patient outcomes.

REFERENCES

[1] M. D. Huffman , D. Prabhakaran , C. Osmond , C. H. D. Fall , . N. Tandon , L. Ramakrishnan, S. Ramji and A. Khalil, “National Library of Medicine,” National Center for Biotechnological Information, 26 Apr 2011. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3408699/.

[2] S. Varughese and G. Abraham, “National Library of Medicine,” National Center for Biotechnological Information, 30 Jan 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5969474/.

[3] L. D. Clin , “National Library of Medicine,” National Center for Biotechnological Information, 28 Jan 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8958241/.

[4] I. J. Ophthalmol, “National Library of Medicine,” National Center for Biotechnological Information, 29 Oct 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8725109/.

[5] “Kaggle,” [Online]. Available: https://www.kaggle.com/.

[6] G. f. Geeks, “Geeks for Geeks,” Geeks for Geeks, 2025. [Online]. Available: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/.

[7] G. f. Geeks, “Geeks for Geeks,” Geeks for Geeks, 2024. [Online]. Available: https://www.geeksforgeeks.org/what-is-feature-extraction/.

[8] G. f. Geeks, “Geeks for Geeks,” Geeks for Geeks, 2024. [Online]. Available: https://www.geeksforgeeks.org/top-6-machine-learning-algorithms-for-classification/.

[9] Franklin, R. G. and B. Muthukumar, “Survey of heart disease prediction and identification using machine learning approaches.,” 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 553-557, 2020.

[10] Sharma, V. S. Yadav and M. Gupta, “Heart disease prediction using machine learning techniques.,” 2nd international conference on advances in computing, communication control and networking (ICACCCN)., pp. 177-181, 2020.

[11] S. Ambesange, A. Vijayalaxmi, S. Sridevi and B. Yashoda, “Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques,” 2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4), pp. 827-832, 2020.

[12] Rani and J. K., “Diabetes prediction using machine learning.,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology 6, no. 4, pp. 294-305, 2020.

[13] Mujumdar, Aishwarya and V. Vaidehi, “Diabetes prediction using machine learning algorithms.,” Procedia Computer Science 165, pp. 292-299, 2019.

[14] Islam, M. Ariful, M. Z. H. Majumder and M. A. Hussein, “Chronic kidney disease prediction based on machine learning algorithms.,” Journal of pathology informatics 14, p. 100189, 2023.

[15] Ekanayake, I. Udara and H. Damayanthi, “Chronic kidney disease prediction using machine learning methods.,” Moratuwa Engineering Research Conference (MERCon), pp. 260-265, 2020.

[16] R. Bagawade, R. Manda, J. gavit, V. Karanjkar and A. Lohar, “HEART DISEASE PREDICTION USING ENHANCED DEEP LEARNING,” International Research Journal of Modernization in Engineering Technology and Science, vol. 06, no. 05, 2024.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING