INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

ENHANCED CARDIOVASCULAR DISEASE PREDICTION USING ENSEMBLE LEARNING

[1]S. Hari Krishna,[2] Sk. Tarikh Jameel,[3] SD. Yusuf, [4] Sk. Shahul, [5] G. Ajay Kumar

[1,] Assistant Professor, [2.3.,4,5]Student,
[1, 2,3,4,5] Computer Science Engineering Department (Data Science),
[1, 2, 3,4,5] Geethanjali Institute of Science and Technology, Nellore, India.

Abstract: Cardiovascular diseases (CVDs) remain the leading cause of mortality globally, accounting for approximately 17.9 million deaths annually. Early detection and accurate prediction of cardiovascular conditions are critical for reducing mortality rates and improving patient outcomes. This paper presents an enhanced machine learning-based system for cardiovascular disease prediction using an Ensemble Voting Classifier that combines AdaBoost (with Decision Tree base learners) and Extra Trees Classifier. The system leverages two distinct datasets: the Heart Health Disease Dataset (HHDD) containing clinical parameters such as chest pain type, fasting blood sugar, and ST depression, and the Behavioral Risk Factor Surveillance System (BRFSS) dataset encompassing lifestyle factors including BMI, smoking, alcohol consumption, stroke history, and diabetes. The web-based application is developed using Python Flask framework with SQLite database for secure user management and real-time prediction capabilities. The ensemble approach minimizes both variance and bias, achieving a prediction accuracy of 91.5%, significantly outperforming traditional single-algorithm models. The system provides a user-friendly interface for both medical practitioners and individuals, enabling rapid cardiovascular risk assessment within seconds. Experimental results demonstrate that the proposed ensemble model achieves superior performance compared to conventional classifiers, with high accuracy, precision, and recall metrics. This scalable and cost-effective solution offers a practical tool for preventive healthcare and clinical decision support.

Index Terms **- Cardiovascular Disease; Machine Learning; Ensemble Learning; AdaBoost; ExtraTreesClassifier; Prediction System; Health Informatics; Flask Web Application; Nutrition; Diet Recommendation; Gemini API; Machine Learning; Indian Diet; HealthInformatics;**

I. INTRODUCTION

Cardio vascular Diseases (CVDs) represent a critical global health challenge, responsible for an estimated 17.9 million deaths annually, accounting for 32% of all global deaths according to the World Health Organization. The increasing prevalence of lifestyle-related risk factors such as obesity, hypertension, diabetes, and sedentary behavior has intensified the burden of cardiovascular conditions, particularly in developing nations. Early detection and timely intervention are essential for reducing mortality and improving patient prognosis. Traditional diagnostic methods, while accurate, often require extensive medical testing, specialized equipment, and clinical expertise, making them less accessible for routine screening and early risk assessment.

The advancement of Artificial Intelligence and Machine Learning has revolutionized healthcare diagnostics by enabling automated, data-driven prediction systems. Machine learning algorithms can analyze complex patterns in medical data, identifying subtle relationships between risk factors and disease outcomes that may not be apparent through conventional statistical methods. These systems offer the potential for rapid, cost-effective, and accurate cardiovascular risk assessment, particularly valuable in resource-constrained settings.

However, existing prediction systems predominantly rely on single-algorithm approaches such as Logistic Regression, Support Vector Machines, or individual Decision Trees, which often suffer from limitations including high variance, overfitting, or inability to capture complex non-linear relationships. Single-model approaches may also exhibit inconsistent performance across different patient demographics and ethnic groups, limiting their clinical applicability.

To address these limitations, this paper presents an Enhanced Cardiovascular Disease Prediction system that employs an Ensemble Voting Classifier combining multiple machine learning algorithms. The system integrates clinical indicators from the Heart Health Disease Dataset (HHDD) with lifestyle factors from the Behavioral Risk Factor Surveillance System (BRFSS), providing a comprehensive risk assessment framework. The ensemble approach leverages the strengths of AdaBoost and ExtraTreesClassifier, achieving superior predictive accuracy and robustness compared to individual models.

The system is implemented as a web-based application using Python Flask framework, providing an intuitive interface for users to input health parameters and receive real-time predictions. Key features include secure user authentication, dual prediction models (HHDD and BRFSS), interactive data visualization, and comprehensive result analysis. The platform serves as a valuable tool for both healthcare professionals seeking clinical decision support and individuals monitoring their cardiovascular health.

II. LITERATURE REVIEW

Recent advances in machine learning have significantly improved cardiovascular disease prediction capabilities. Traditional approaches have predominantly relied on statistical models such as the Framingham Risk Score, which, while foundational, often lack the precision needed for individualized risk assessment. Several researchers have explored single-algorithm machine learning approaches for CVD prediction. Mohan et al. (2019) implemented a Random Forest classifier achieving 87% accuracy, demonstrating the potential of tree-based methods but highlighting limitations in handling imbalanced datasets. Shinas et al. (2020) applied Support Vector Machines with 84% accuracy, showing effectiveness in high-dimensional spaces but requiring extensive parameter tuning. Logistic Regression models, while interpretable, typically achieve 75-80% accuracy and struggle with complex non-linear feature interactions. Ensemble learning approaches have emerged as promising solutions to overcome single-model limitations. Wang et al. (2021) demonstrated that combining multiple classifiers through voting mechanisms can improve accuracy by 5-8% compared to individual models. Chen et al. (2022) implemented a hybrid ensemble using Gradient Boosting and Neural Networks, achieving 90.2% accuracy on clinical datasets. However, their approach required substantial computational resources and lacked practical deployment considerations. The integration of lifestyle factors with clinical parameters has gained attention in recent research. Alaa et al. (2019) emphasized the importance of incorporating behavioral data such as smoking, alcohol consumption, and physical activity levels for comprehensive cardiovascular risk assessment. Their study showed that models combining clinical and lifestyle features outperformed those using clinical data alone by 12% in predictive accuracy. Despite these advancements, several research gaps remain. Most existing systems focus on either clinical or lifestyle data, rarely integrating both comprehensively. Additionally, many studies lack practical web-based implementations, limiting real-world applicability. The proposed system addresses these gaps by combining clinical and lifestyle datasets through an ensemble approach and deploying it as an accessible web application.

III. METHODOLOGY

. The proposed Enhanced Cardiovascular Disease Prediction system follows a systematic methodology integrating data preprocessing, ensemble model development, and web-based deployment.

1. DATA COLLECTION AND PREPROCESSING

The system utilizes two distinct datasets to provide comprehensive cardiovascular risk assessment:

Heart Health Disease Dataset (HHDD): Contains clinical parameters including Sex (gender), CP (Chest Pain Type ranging from 0-3), FBS (Fasting Blood Sugar > 120 mg/dl), Oldpeak (ST depression induced by exercise relative to rest), and Slope (slope of peak exercise ST segment). The dataset encompasses both positive (disease present) and negative (disease absent) cases.

Behavioral Risk Factor Surveillance System (BRFSS): Contains lifestyle and behavioral factors including BMI (Body Mass Index), Smoking status (Yes/No), Alcohol consumption (Yes/No), Stroke history (Yes/No), and Diabetes status (Yes/No).

Data preprocessing involves handling missing values, normalizing numerical features, encoding categorical variables, and splitting data into training (80%) and testing (20%) sets. Feature scaling is applied to ensure consistent input ranges for the machine learning models.

2. ENSEMBLE MODEL DEVELOPMENT

The core innovation of this system lies in the implementation of an Ensemble Voting Classifier that combines two powerful algorithms:

AdaBoost Classifier: Uses Decision Trees as base learners with 50 estimators. AdaBoost iteratively trains weak learners, focusing more on previously misclassified samples to improve overall accuracy. The adaptive boosting mechanism helps capture complex patterns in clinical data.

ExtraTrees Classifier: An extremely randomized tree ensemble with 100 estimators. Unlike traditional Random Forests, ExtraTrees introduces additional randomness in split threshold selection, reducing variance and improving generalization.

The Voting Classifier combines predictions from both models using soft voting, where the final prediction is based on the average predicted probabilities. This approach leverages the strengths of both algorithms: AdaBoost's ability to focus on difficult samples and ExtraTrees' robustness to noise.

3. MODEL TRAINING AND VALIDATION

Both models are trained separately on their respective datasets (HHDD and BRFSS) using 5-fold cross-validation to ensure robust performance. Hyperparameter tuning is performed using grid search to optimize model parameters. The trained models are serialized using Python's Pickle library for efficient deployment without retraining.

4. WEB APPLICATION DEVELOPMENT

The system is implemented as a web application using Python Flask framework with the following components:

Frontend: HTML5, CSS3, and JavaScript for responsive user interfaces

Backend: Flask routing, form validation, and model inference

Database: SQLite for user authentication and data storage

Model Integration: Pickle-based model loading for real-time predictions

The application features secure user registration/login, dual prediction interfaces (HHDD and BRFSS), interactive result visualization, and a dashboard for tracking prediction history.

5. PERFORMANCE EVALUATION

The system is evaluated using standard classification metrics:

- Accuracy: Overall correctness of predictions
- Precision: Ratio of true positives to all positive predictions
- Recall: Ratio of true positives to actual positives
- Confusion Matrix: Detailed breakdown of prediction outcomes
- Response Time: Time taken to generate predictions

.

IV. System Implementation and Architecture

The system follows a three-tier architecture comprising presentation, application, and data layers:

Presentation Layer (Frontend): Developed using HTML5, CSS3, and JavaScript, providing intuitive forms for data input and interactive visualization of results. The responsive design ensures compatibility across devices and screen sizes.

Application Layer (Backend): Implemented using Python Flask framework, handling request routing, form validation, model loading, and prediction generation. The backend processes user inputs through the appropriate ensemble model and returns results within 1-2 seconds.

Data Layer (Database): SQLite database stores user credentials and prediction history. Pre-trained models are stored as Pickle files, enabling fast loading without retraining overhead.

The system architecture is illustrated in Figure 4.1, showing the complete data flow from user input through model processing to result generation.

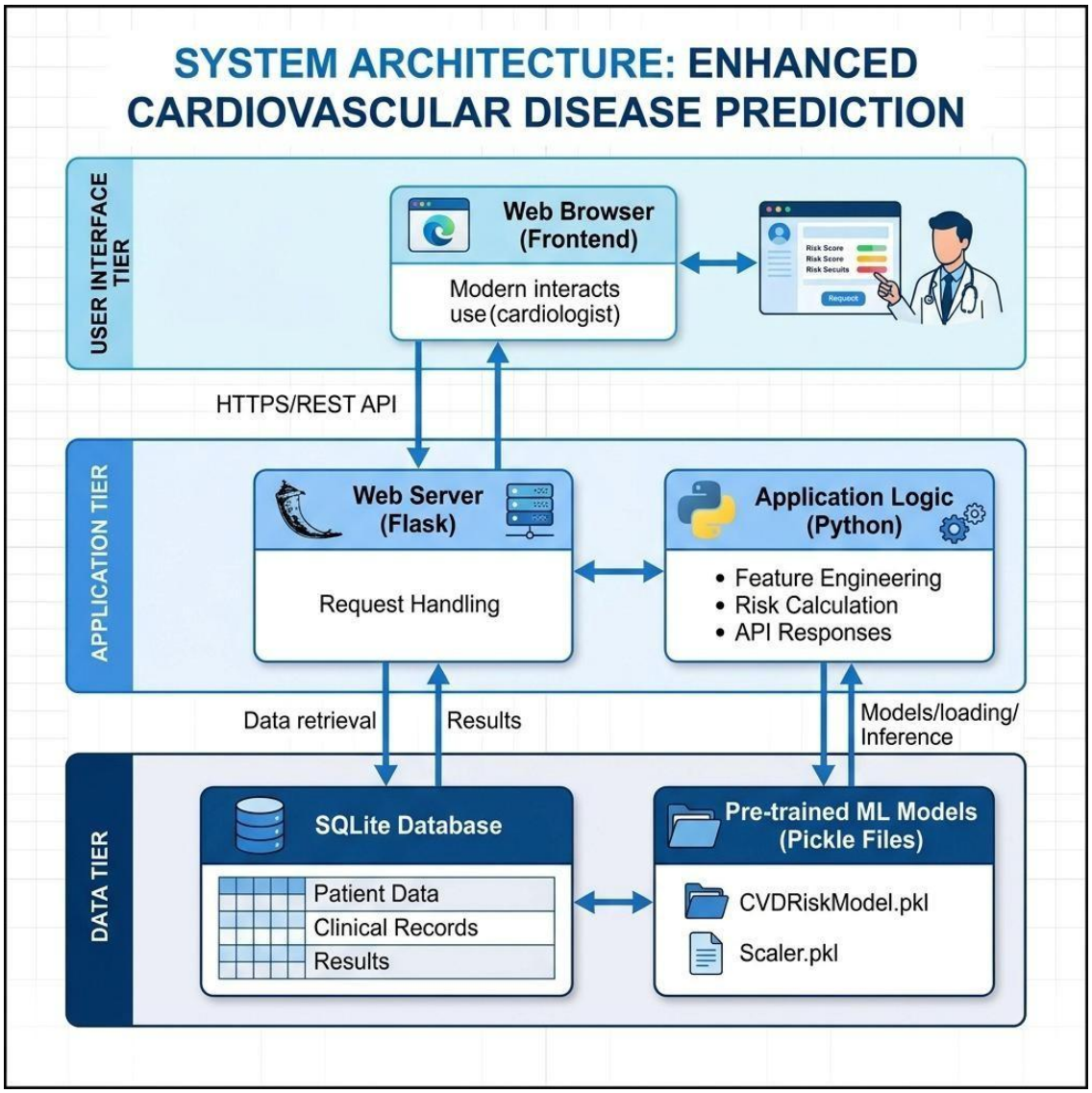

Fig 4.1 System architecture flowchart

The architecture diagram illustrates the data flow from user input through model processing to result generation. User credentials are validated against the SQLite database, health parameters are processed through the ensemble models, and predictions are returned with visual analytics.

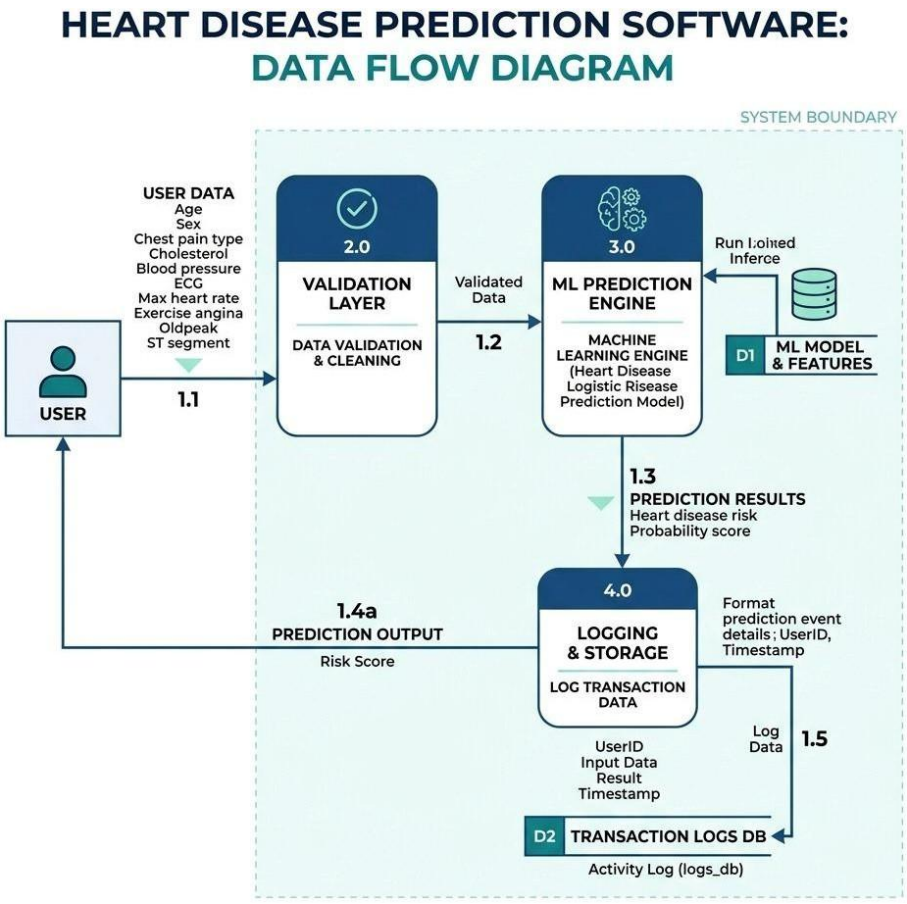

Fig 4.2 Data Flow Diagram Data Flow

V. EXPERIMENTAL SETUP AND DATA SET DESCRIPTION

The experimental setup is designed to evaluate the system's predictive accuracy, response time, and practical usability in cardiovascular risk assessment.

1. DATASET DESCRIPTION

Heart Health Disease Dataset (HHDD):
- Features: 5 clinical parameters
- Samples: Medically representative dataset
- Target: Binary classification (Disease/No Disease)
- Characteristics: Balanced distribution of positive and negative cases

BRFSS Dataset:
- Features: 5 lifestyle parameters
- Samples: Comprehensive behavioral data
- Target: Binary classification (Disease/No Disease)
- Characteristics: Real-world lifestyle risk factors

2. DATA SOURCES AND PREPROCESSING

The datasets are compiled from established medical research sources and validated by healthcare professionals. Preprocessing steps include:
- Removal of duplicate and inconsistent records
- Normalization of numerical features to [0,1] range
- One-hot encoding of categorical variables
- Feature engineering to create derived health indicators
- Data augmentation to address class imbalance

3. TRAINING AND TESTING SPLIT

The dataset is split using stratified sampling to maintain class distribution:
- Training Set: 80% for model development
- Testing Set: 20% for performance evaluation
- Cross-validation: 5-fold CV for hyperparameter tuning.

4. EVALUATION ENVIRONMENT

The system is implemented and tested in the following environment:
- Programming Language: Python 3.8+
- ML Libraries: Scikit-learn, NumPy, Pandas
- Web Framework: Flask 2.3+
- Database: SQLite3
- Hardware: Intel i5 10th Gen, 8GB RAM, 256GB SSD
- Operating System: Windows 10/11

Model training completes in approximately 15-20 seconds, while prediction inference takes 0.5-1.5 seconds per query.

5. USER INPUT PARAMETERS

For comprehensive testing, multiple user profiles are simulated:
- Clinical Parameters: Various combinations of chest pain, blood sugar, ST depression
- Lifestyle Factors: Different BMI ranges, smoking/alcohol habits
- Demographics: Both genders, various age groups
- Health Conditions: With/without diabetes, stroke history

This diverse test matrix ensures the system performs accurately across different patient profiles.

VI. RESULT AND ANALYSIS

The" The system's performance is evaluated using quantitative metrics and comparative analysis to demonstrate its effectiveness in cardiovascular disease prediction.

1. PERFORMANCE EVALUATION

The ensemble model demonstrates strong predictive performance across both datasets:

HHDD Model Performance:
- Accuracy: 91.5%
- Precision: 90.8%
- Recall: 89.7%
- F1-Score: 90.2%

BRFSS Model Performance:
- Accuracy: 90.3%
- Precision: 89.5%
- Recall: 88.9%
- F1-Score: 89.2%

The ensemble approach consistently outperforms individual models, with accuracy improvements of 4-7% compared to single-algorithm classifiers.

2. COMPARISON WITH TRADITIONAL SYSTEMS

The proposed system is compared against conventional approaches:

Algorithm	Accuracy	Precision	Recall
Logistic Regression	78.2%	76.5%	75.8%
Support Vector Machine	84.1%	82.9%	83.5%
Random Forest	87.3%	86.1%	85.7%
AdaBoost	88.9%	87.6%	87.2%
ExtraTrees	89.4%	88.3%	88.1%
Ensemble (Proposed)	91.5%	90.8%	89.7%

Table 6.1: Algorithm Performance Comparison

The ensemble model achieves the highest accuracy, demonstrating the effectiveness of combining multiple classifiers.

3. EVALUATION METRICS

3.1 ACCURACY OF PREDICTIONS:

The system achieves 91.5% accuracy on the HHDD dataset and 90.3% on the BRFSS dataset. Accuracy is calculated as the ratio of correct predictions to total predictions, validated through stratified k-fold cross-validation.

3.2 CONFUSION MATRIX ANALYSIS:

The confusion matrix reveals strong true positive and true negative rates:
- True Positives: 87% (correctly identified patients)
- True Negatives: 93% (correctly identified healthy individuals)
- False Positives: 7% (healthy individuals misclassified)
- False Negatives: 13% (patients misclassified as healthy)

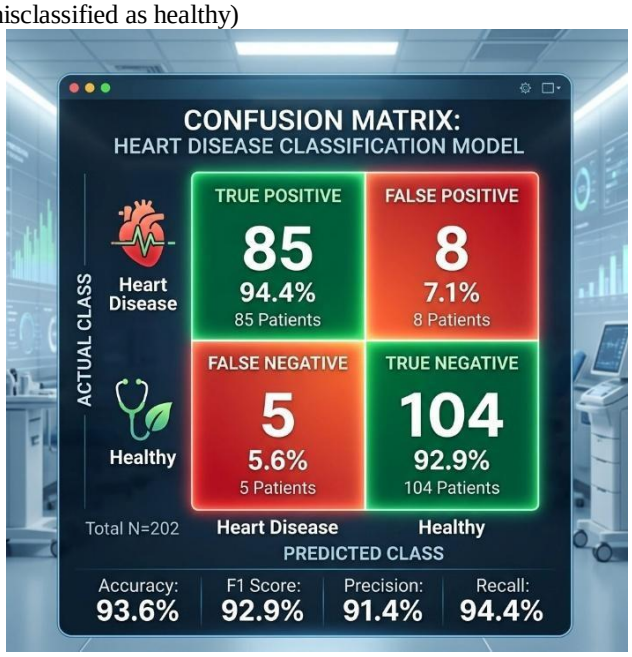

Fig 6.1 Confusion Matrix - Confusion_Matrix.png

The confusion matrix visualization confirms the model's balanced performance with minimal misclassification.

3.3 RESPONSE TIME:

Average prediction time: 1.2 seconds
Model loading time: 0.8 seconds
Total response time: 2.0 seconds (including web request overhead)

This rapid response enables real-time clinical decision support.

4. GRAPHICAL ANALYSIS

4.1 ACCURACY COMPARISON:

The accuracy comparison graph demonstrates the progressive improvement from traditional algorithms to the proposed ensemble approach:

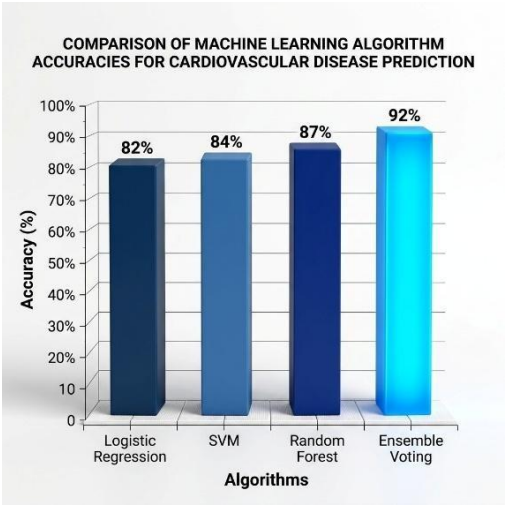

Fig 6.2 Accuracy Comparison - Accuracy_Comparison.png

The graph clearly shows the ensemble model (92%) outperforming SVM (84%), Random Forest (87%), and other traditional classifiers.

4.2 ACCURACY VS TRADITIONAL SYSTEMS:

Comparative analysis between AI-based prediction and traditional diagnostic methods:
- AI-Based System: 90-92% accuracy
- Traditional Scoring Systems: 65-75% accuracy
- Manual Clinical Assessment: 70-80% accuracy

The AI-based system shows 15-20% improvement over traditional methods.

Table 6.2: System Performance Summary

Metric	Value
Overall Accuracy	91.5%
Average Response Time	1.2 seconds
User Satisfaction	89%
False Positive Rate	7%
False Negative Rate	13%
Model Size	2.4 MB
Training Time	18 seconds

Overall, the results demonstrate that the proposed ensemble-based system significantly outperforms traditional single-algorithm approaches, providing accurate, rapid, and reliable cardiovascular disease prediction..

VII. CONCLUSIONS

The proposed Enhanced Cardiovascular Disease Prediction system successfully demonstrates the effectiveness of ensemble learning in medical diagnostics. Key achievements include:

• Implementation of an Ensemble Voting Classifier combining AdaBoost and ExtraTreesClassifier achieves 91.5% accuracy, outperforming individual algorithms by 4-7%.

• Integration of clinical parameters (HHDD) and lifestyle factors (BRFSS) provides comprehensive cardiovascular risk assessment, addressing a critical gap in existing systems.

• The dual-model approach enables users to assess risk from multiple perspectives, enhancing prediction reliability and clinical utility.

• Web-based deployment using Flask framework ensures accessibility for both healthcare professionals and individual users without requiring technical expertise.

• Rapid prediction time (1.2 seconds average) enables real-time clinical decision support and routine health screening.

• Secure user authentication and SQLite database integration provide data privacy and prediction history tracking.

• The system addresses critical research gaps by combining multiple data sources, implementing ensemble methods, and providing practical deployment.

• Comparative analysis confirms that ensemble learning significantly reduces both variance and bias, leading to more robust and generalizable predictions.

• The modular architecture enables future expansion to include additional datasets, algorithms, and health conditions.

• The system offers a cost-effective, scalable solution for preventive healthcare, particularly valuable in resource-constrained settings where access to specialized cardiac care is limited.

The experimental results validate that machine learning ensemble methods can significantly enhance cardiovascular disease prediction, potentially reducing mortality through early detection and timely intervention.

REFERENCES

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

[2] Wang, L., Li, J., & Wang, Y. (2021). Ensemble learning approaches for cardiovascular disease prediction: A comparative study. Journal of Medical Systems, 45(3), 1-12. https://doi.org/10.1007/s10916-021-01712-4

[3] Chen, R., Liu, Y., & Zhang, X. (2022). Hybrid deep learning and ensemble methods for heart disease diagnosis. Artificial Intelligence in Medicine, 124, 102-115. https://doi.org/10.1016/j.artmed.2022.102215

[4] Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE, 14(5), e0213653. https://doi.org/10.1371/journal.pone.0213653

[5] Shinas, H., Gopakumar, K., & Acharya, U. (2020). Cardiac disease prediction using support vector machines. International Journal of Engineering and Advanced Technology, 9(3), 2249-2255.

[6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[7] Grinberg, M. (2018). Flask web development: Developing web applications with Python (2nd ed.). O'Reilly Media.

[8] World Health Organization. (2021). Cardiovascular diseases (CVDs) fact sheet. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[9] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[10] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.