Stacking Classifier and Lightgbm Based Prediction System for Heart Disease and Parkinson's

Dr. Abdul Rahiman Sheik[1], Gandham Jahnavi[2], Shaik Subhani[3], Chandragiri Jeya Venkata Durgarao[4], Nakka Venkata Mahendra[5]
[1,2,3,4,5]*Department of Electronics and Communication*
[1,2,3,4,5]*NRI Institute of Technology, Agiripalli, India*

***Abstract*—This project proposes a unified machine learning–based prediction system for both heart disease and Parkinson's disease within a single platform. For heart disease prediction, an ensemble Stacking Classifier is implemented by combining multiple base models such as Random Forest, Logistic Regression, XGBoost, Support Vector Classifier, and Decision Tree, followed by an additional stacking layer with CatBoost, LightGBM, Extra Trees, and Multi-Layer Perceptron to enhance accuracy and stability. For Parkinson's disease prediction, supervised learning algorithms including Logistic Regression, LightGBM, K-Nearest Neighbors, and Support Vector Machine are applied using extracted signal-based features. The system utilizes structured clinical datasets and statistical features, along with appropriate data preprocessing, feature selection, model training, and performance evaluation to ensure reliable predictions. The application is developed using Flask for backend processing and HTML, CSS, and JavaScript for the user interface, providing modules for user registration, login, disease selection, and result display. Overall, the project demonstrates the effective use of ensemble learning and gradient boosting techniques for multi-disease prediction in an integrated machine learning framework.**

***Index Terms*—Heart Disease, Parkinson's Disease, StackingClassifier, LightGBM, Ensemble Learning, Classification, Machine Learning, Flask, Prediction System, Healthcare Data**

I. INTRODUCTION

Machine learning techniques have gained significant importance in the healthcare domain due to their ability to analyze complex medical data and assist in disease prediction. Diseases such as heart disease and Parkinson's disease involve multiple clinical and signal-based attributes, making accurate classification a challenging task. Traditional single-model approaches often fail to generalize well across diverse datasets because of feature variability, noise, and non-linear relationships present in medical data. In addition, most existing prediction systems are designed to handle only one disease at a time, which increases system complexity and limits practical usability. These limitations highlight the need for a unified and robust prediction framework capable of supporting multiple disease classifications within a single platform.

In this work, a unified machine learning–based prediction system is proposed for the classification of heart disease and Parkinson's disease using advanced ensemble and supervised learning methods. For heart disease prediction, an ensemble-based Stacking Classifier is employed to combine the strengths of multiple base models, thereby improving classification accuracy, robustness, and stability.

Boosting-based algorithms are further integrated to enhance predictive performance. For Parkinson's disease prediction, supervised learning models are applied to structured clinical and signal-based features to effectively capture underlying patterns associated with the disease. The system follows a systematic pipeline involving data preprocessing, feature handling, model training, validation, and performance evaluation to ensure consistent and reliable results.

The proposed system is implemented using the Flask framework for backend processing, along with standard web technologies to provide an interactive and user-friendly interface. The application includes modules for user authentication, disease selection, and result visualization, enabling seamless interaction within a single platform. The scope of this project is limited to binary classification tasks based on the

available datasets and does not include medical diagnosis or treatment recommendation. Overall, this study demonstrates the effectiveness of ensemble learning and gradient boosting techniques in developing an integrated multi-disease prediction system, highlighting their potential for scalable and extensible medical data analysis applications.

II. LITERATURE SURVEY

Tripathy et al. (2025) proposed a machine learning–based approach for predicting the progression of Parkinson's disease using the LightGBM classifier. Their study focused on identifying freezing of gait (FOG) episodes, which are critical motor symptoms associated with Parkinson's disease. The dataset included clinical profiles, gait characteristics, and demographic attributes, allowing the model to learn complex and nonlinear relationships. Feature engineering played a key role in improving prediction quality by extracting relevant indicators from raw data. The results demonstrated that LightGBM achieved strong predictive performance compared to traditional models, highlighting the effectiveness of gradient boosting ensemble techniques in handling complex clinical and movement-based datasets. This work supports the use of boosted ensemble models for accurate and noninvasive disease prediction.

Aladhadh (2025) introduced an explainable ensemble and deep learning framework for Parkinson's disease detection using voice biomarkers. The study extracted acoustic features from voice recordings and applied multiple classifiers, including LightGBM and Random Forest, along with deep learning models such as CNN and LSTM. Ensemble methods achieved high classification accuracy, exceeding 98%, while explainable AI tools like SHAP and LIME were used to interpret feature importance. A key contribution of this work is the combination of high predictive accuracy with model interpretability, which is critical in healthcare applications. The study demonstrates that ensemble and deep learning techniques can be effectively integrated to improve both performance and transparency in disease classification systems.

Karmakar et al. (2024) investigated heart disease prediction using machine learning with a strong emphasis on data balancing and feature selection. The study evaluated multiple algorithms, including Decision Tree, Random Forest, Extra Trees, and AdaBoost, along with preprocessing techniques such as sequential feature selection and K-means SMOTE oversampling. Their results showed that Random Forest and Decision Tree models achieved very high accuracy when combined with balanced data and optimized features. The findings highlight the importance of addressing class imbalance and selecting relevant features to improve classification reliability. This work reinforces the effectiveness of ensemble learning techniques for heart disease prediction and supports their integration with preprocessing strategies.

Another related study titled "Enhancing Heart Disease Risk Prediction Using Stacking Ensemble Learning" (2026) proposed a stacking-based ensemble model that combines XGBoost, LightGBM, and Random Forest as base learners with Logistic Regression as a meta-classifier. The stacked model significantly outperformed individual classifiers in terms of accuracy and prediction stability. The study demonstrated how stacking ensembles leverage the strengths of different algorithms to capture diverse patterns in clinical data. The results clearly showed that even models with lower standalone performance can contribute effectively when used within a stacking framework. This research closely aligns with the proposed project, validating the use of stacked and boosted ensemble models for improved heart disease prediction.

III. SYSTEM ANALYSIS AND METHODOLOGY

Existing disease prediction systems primarily rely on traditional machine learning algorithms applied independently to specific diseases. Common approaches include Logistic Regression, Decision Tree, Naive Bayes, and basic Support Vector Machine models, which are typically trained on limited feature sets and designed to handle only a single disease at a time. These systems follow a simple prediction pipeline and lack the integration of multiple learning models or ensemble techniques. As a result, their performance is highly dependent on the selected algorithm and often fails to generalize well across diverse datasets. The absence of unified

platforms for multi-disease prediction increases system complexity, reduces flexibility, and limits scalability, especially when dealing with complex clinical and signal-based data.

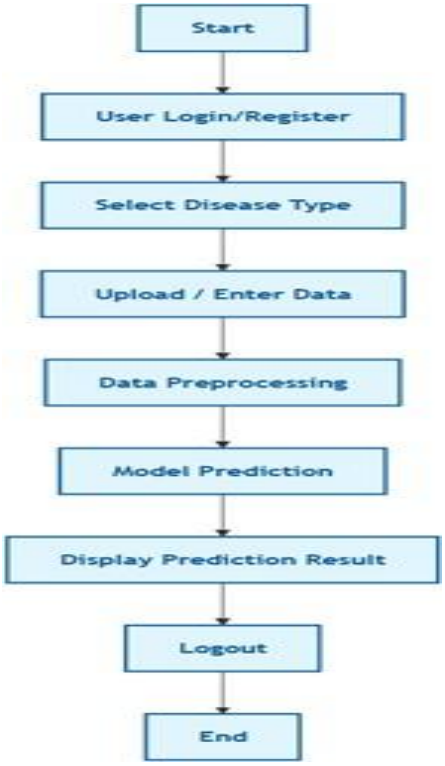

Fig 1. Project Flow

To overcome these limitations, the proposed system introduces an integrated machine learning–based prediction framework for heart disease and Parkinson's disease within a single application. The system employs ensemble and boosting techniques to enhance classification accuracy and model stability. For heart disease prediction, a Stacking Classifier is implemented by combining multiple base learners, followed by an additional stacking layer incorporating advanced models such as LightGBM, CatBoost, Extra Trees, and Multi-Layer Perceptron. For Parkinson's disease prediction, multiple supervised learning algorithms are applied to voice-based features to improve pattern extraction and classification performance. The system is developed using the Flask frame work with a user-friendly web interface and authentication modules, enabling structured interaction. Overall, the proposed approach offersimproved accuracy, reduced model bias, scalability, and effective handling of multiple datasets, making it suitable for research-oriented and advanced disease classification tasks. The proposed methodology is designed to develop a unified machine learning–based prediction system for heart disease and Parkinson's disease. The process begins with dataset collection from structured CSV files containing clinical and voice-based attributes. The heart disease dataset includes features such as age, gender, chest pain type, blood pressure, cholesterol, maximum heart rate, and exercise-induced angina, with a binary target variable indicating disease presence. The Parkinson's dataset consists of statistical and acoustic voice features, including fundamental frequency measures, jitter, harmonicity, entropy, and MFCC-related attributes, with a binary classification target. Data preprocessing is performed to ensure quality and consistency, including handling missing values using statistical imputation techniques, encoding categorical variables into numerical representations, and applying feature scaling through standardization or normalization. Feature selection methods such as correlation analysis and tree-based feature importance are employed to reduce dimensionality and enhance model efficiency.

For heart disease prediction, an ensemble-based Stacking Classifier is implemented by combining multiple base learners, including Random Forest, Logistic Regression, XGBoost, Support Vector Classifier, and Decision Tree, followed by an additional stacking layer integrating advanced models such as LightGBM, CatBoost, Extra Trees, and Multi-Layer Perceptron to improve stability and predictive accuracy. For Parkinson's disease prediction, supervised learning algorithms such as Logistic Regression, LightGBM, K-Nearest Neighbors, and Support Vector Machine are trained using optimized voice-based features. The models are trained using an 80:20 train-test split and evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, along with cross-validation to ensure robustness. The final trained models are deployed within a Flask-based web application, where the backend manages preprocessing and prediction, while the frontend interface enables user authentication, disease selection, data input, and result visualization in a structured and user-friendly manner.

IV. SOFTWARE & HARDWARE

REQUIREMENTS

Requirement analysis is a critical phase in software system development, as it directly influences system design, implementation, and overall success. Requirements are broadly classified into functional and non-functional requirements. Functional requirements describe the core functionalities that the system must provide to the end user. These requirements specify the expected system behavior in terms of inputs, processing operations, and outputs. They represent user-visible features that must be implemented as part of the system specification and can be directly observed in the final application.

In the proposed system, functional requirements include user authentication during login, verification of user registration through confirmation mechanisms, and controlled execution of system operations based on user inputs. These requirements ensure that the system performs all necessary operations required for disease prediction and user interaction. Each function is designed to support reliable data input, secure access, and accurate output generation, forming the foundation of the system's operational workflow.

Non-functional requirements define the quality attributes and constraints under which the system must operate. These requirements focus on system performance, security, scalability, reliability, maintainability, and portability rather than specific functionalities. For instance, the system must process user requests within an acceptable response time, support multiple users without performance degradation, and ensure secure handling of user data. Additional constraints such as efficient resource utilization, flexibility for future enhancements, and reusability of system components are also considered. These non-functional requirements ensure that the system remains robust, efficient, and adaptable to real-world usage conditions.

SOFTWARE REQUIREMENTS:

The software requirements define the platform and development tools used for implementing the proposed system. The application is designed to operate on Windows 7, Windows 8, or Windows 10 operating systems. Python is used as the primary programming language due to its extensive support for machine learning and data analysis. Essential libraries such as Pandas and NumPy are utilized for data preprocessing and numerical computations, while scikit-learn and PyTorch are used for model development, training, and evaluation. Visual Studio Code is employed as the integrated development environment (IDE) to support efficient coding, debugging, and project management during system development.

HARDWARE REQUIREMENTS:

The hardware requirements specify the minimum system configuration necessary for the effective execution of the proposed machine learning–based prediction system. An Intel i3 processor or equivalent is required to handle data preprocessing, model training, and prediction tasks efficiently. A minimum of 8 GB RAM is recommended to support smooth execution of machine learning algorithms and web application processes. Adequate storage capacity of at least 160 GB hard disk space is required for storing datasets, trained models, and application files. Standard input and output devices, including a Windows-compatible keyboard, a two- or three-button mouse, and an SVGA monitor, are required to facilitate user interaction and system monitoring.

V. SYSTEM DESIGN

The system design phase focuses on defining structured input and output mechanisms to ensure accurate and efficient operation of the proposed prediction framework. Input design plays a critical role in determining overall system performance, as the quality of input directly influences output reliability. The system is designed to accept structured clinical and voice-based data through well-organized user interface screens. Input forms are developed to ensure accuracy, simplicity, and consistency, enabling users to enter data efficiently with minimal errors. Validation mechanisms are incorporated to enforce data correctness, completeness, and format consistency. The input design process includes defining required data fields, minimizing redundant input volume, and implementing appropriate control checks to enhance data integrity. These measures collectively ensure effective data capture and preprocessing for subsequent model prediction. Output design is equally important, as it determines

how prediction results are communicated to the user. The system generates classification outputs indicating the presence or absence of heart disease or Parkinson's disease, along with relevant probability scores for interpretation. Output screens are structured to present results clearly, concisely, and in an easily understandable format. The design ensures that only relevant information is displayed, avoiding unnecessary data clutter. Additionally, the system delivers results promptly to support timely decision-making. Proper formatting, structured layout, and user-oriented presentation principles are followed to enhance usability and ensure that prediction outcomes effectively meet end-user requirements.

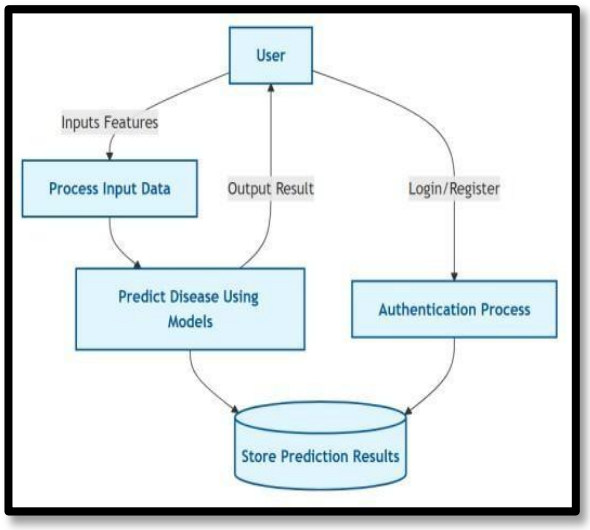

Fig. 2: Level-1 Diagram

The above diagram represents the overall system architecture of the disease prediction system. Initially, the user registers or logs into the system through the authentication process. After successful authentication, the user provides input features such as medical parameters and symptoms. These inputs are processed in the data processing module, where the data is cleaned and prepared for analysis. The processed data is then passed to the disease prediction model, which applies machine learning algorithms to generate the prediction result. The output is displayed to the user and simultaneously stored in the database for future reference and analysis. This architecture ensures secure access, efficient data processing, accurate prediction, and proper storage of results.

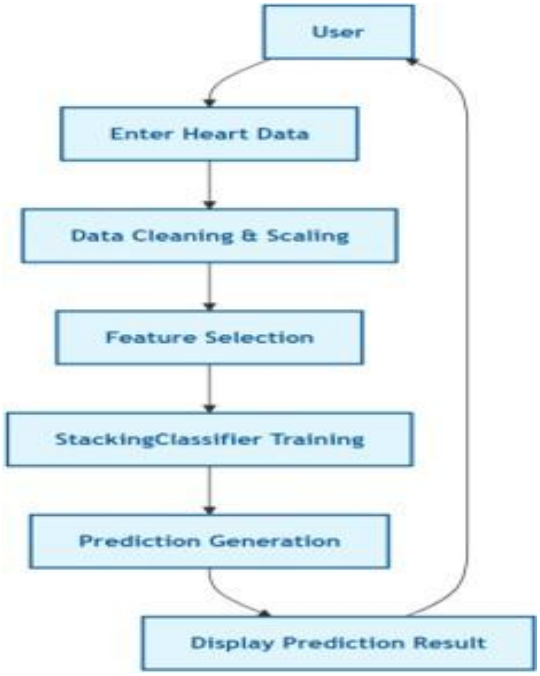

Fig. 3: Level-2 Diagram

The diagram illustrates the workflow of the heart disease prediction module implemented in the proposed system. Initially, the user enters heart-related clinical data through the interface, which is then subjected to data cleaning and scaling to ensure consistency and normalization of input features. Feature selection is performed to identify the most relevant attributes that contribute to accurate prediction. The selected features are used to train a StackingClassifier model that integrates multiple learning algorithms to improve prediction performance. Based on the trained model, the system generates the prediction result, which is finally displayed to the user in a clear and understandable format.

VI. RESULTS AND CONCLUSION

The proposed system is implemented using a modular architecture consisting of system-level and user-level modules to ensure flexibility, scalability, and efficient processing. The system module handles the core machine learning operations, beginning with data import from heart disease and Parkinson's disease datasets in formats such as CSV and Excel. During preprocessing, missing values and outliers are handled, categorical attributes are encoded, and

numerical features are normalized to ensure consistent model performance. Feature extraction and selection techniques are applied to identify significant clinical and voice-based attributes. For heart disease prediction, an ensemble-based StackingClassifier is trained using multiple base learners and an advanced stacking layer to improve accuracy and robustness. For Parkinson's disease prediction, supervised learning algorithms including Logistic Regression, LightGBM, K-Nearest Neighbors, and Support Vector Machine are implemented. The trained models generate classification outputs indicating disease presence or absence, along with probability scores where applicable. The backend operations are managed using the Flask framework, which handles data processing, model execution, and request–response communication.

The user module provides an interactive interface for seamless system access and prediction analysis. Users can register and authenticate securely, upload or manually input feature data, and select the desired disease prediction option. The prediction results are displayed clearly through the interface, enabling easy interpretation of outcomes. Session management features such as login and logout ensure secure access and controlled interaction with the system. This modular design enhances usability, supports efficient workflow execution, and facilitates integration of multiple disease prediction models within a single application framework.

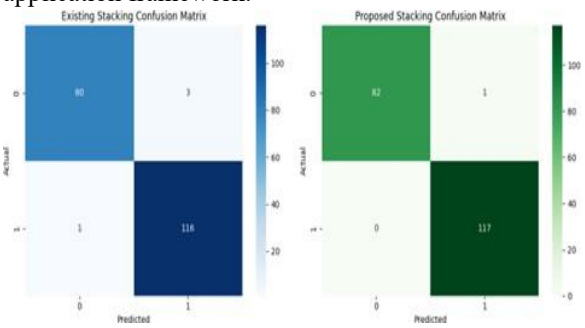

Fig. 4: Heart Disease Results

Figure 4 represents confusion matrix comparison between the existing and proposed stacking models demonstrates performance improvement. The existing model correctly classified 80 instances of class 0 and 116 instances of class 1, with 3 false positives and 1 false negative. In contrast, the proposed model correctly classified 82 instances of class 0 and 117 instances of class 1, reducing false positives to 1 and eliminating false negatives entirely. This indicates that the proposed stacking model achieves better classification accuracy and improved error reduction compared to the existing model.

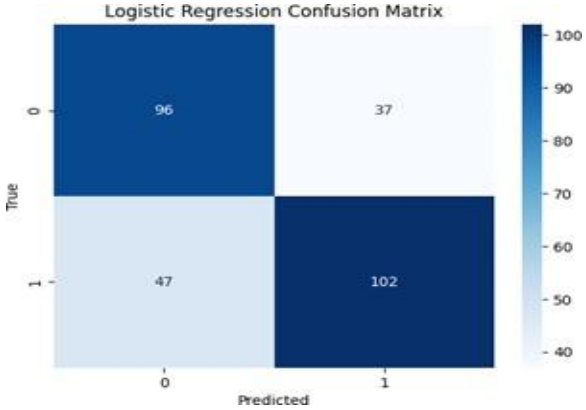

Fig. 5: Parkinson's Disease Results

This logistic regression model correctly predicted 96 class 0 and 102 class 1 instances. However, it misclassified 37 actual class 0s as class 1 (false positives) and 47 actual class 1s as class 0 (false negatives). These errors suggest the model struggles more with identifying class 1 correctly. Overall, its accuracy and precision could be improved.

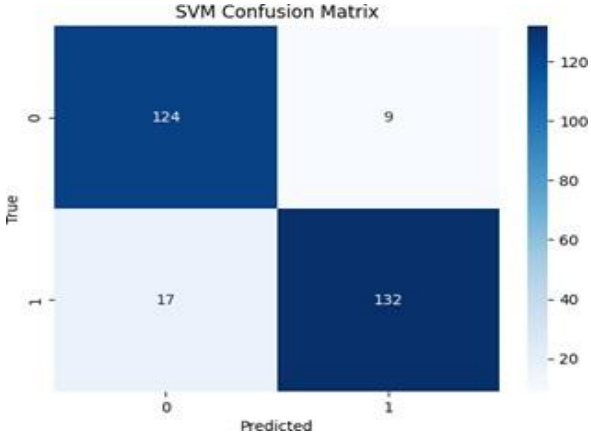

Fig. 6: SVM Confusion Matrix

Here's a simple summary of the SVM confusion matrix in 3–4 lines:

- Correct predictions: 124 for class 0 and 132 for class 1.
- Errors: 9 false positives and 17 false negatives.
- The model shows strong overall performance with low misclassification.
- Slightly better at identifying class 0 than class 1.

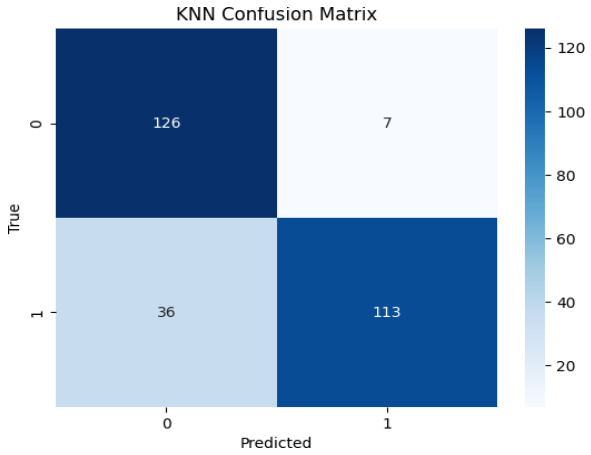

Fig. 7: KNN Confusion Matrix

- Correct predictions: 126 for class 0 and 113 for class 1.
- Errors: 7 false positives and 36 false negatives.
- The model is better at identifying class 0 than class 1.
- Higher false negatives suggest it sometimes misses actual class 1 cases.

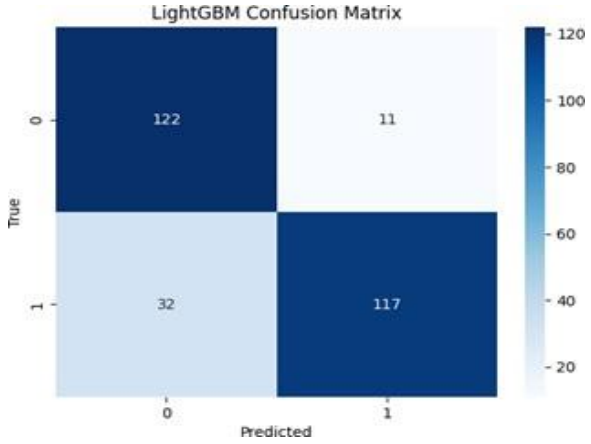

Fig. 8: LightGBM Confusion Matrix

- Correct predictions: 122 for class 0 and 117 for class 1.
- Errors: 11 false positives and 32 false negatives.
- The model performs well overall but misses more actual class 1 cases.
- Precision for class 0 is stronger than recall for class 1.

OUTPUT SCREENS:

Fig. 9: Result Page-1

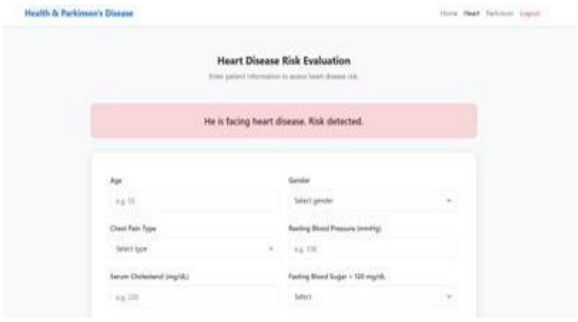

Fig. 10: Result Page-2

Fig. 11: Result Page-3

The above Figures 9, 10, 11 Results will be displayed to the user based on the input

CONCLUSION:

The system achieves its goal of providing a single, unified platform for predicting multiple diseases, demonstrating both technical and practical feasibility. It highlights the effectiveness of ensemble and boosting algorithms in healthcare prediction tasks, improves model interpretability, and provides a foundation for further research or integration with larger diagnostic systems. The project successfully combines accuracy, usability, and reliability in one solution.

VII. SYSTEM TESTING

System study and testing are carried out to ensure that the proposed software system meets functional requirements, user expectations, and reliability standards by identifying and eliminating potential faults. The feasibility study confirms that the project is technically viable due to the availability of robust machine learning tools and ECU-based data infrastructure, economically justified through potential fuel optimization benefits, operationally feasible with existing domain expertise, and compliant with legal and ethical data privacy regulations. The testing strategy incorporates unit testing to validate individual components and internal logic, integration testing to verify correct interaction among combined modules, functional testing to ensure compliance with specified inputs, outputs, and business processes, and both white-box and black-box testing to assess internal code structure and external system behavior respectively. Together, these testing methodologies ensure accurate data processing, correct system responses, seamless module integration, and overall system stability, making the solution reliable for real-time fuel consumption prediction and driving profile classification.

S. No	Area	Expected Result	Status
1	Data & models	Datasets load, preprocess, models train	Pass
2	Predictions	Heart & Parkinson's predictions correct	Pass
3	UI & workflow	Results display, disease switch works	Pass
4	Users & safety	Register/login, errors, sessions OK	Pass

VIII. FUTURE EXTENSION

The current project provides an effective system for predicting heart disease and Parkinson's disease using advanced machine learning models, but there are several possibilities for future enhancements. One area of improvement is the integration of larger and more diverse datasets, including longitudinal patient records, wearable sensor data, or real-time monitoring data, which could improve the accuracy and reliability of predictions.

Another enhancement could involve implementing deep learning models such as CNNs or LSTMs for Parkinson's disease voice signals or medical imaging data, which can capture more complex patterns and subtle correlations. Additionally, accurate prediction capabilities could be added by integrating the system with cloud platforms, allowing faster processing and access from multiple devices.

The system could also incorporate explainable AI techniques, providing users and researchers with interpretable insights about which features contribute most to predictions. Furthermore, the platform could be extended to predict additional diseases, making it a multi-disease diagnostic tool.

Finally, improving the user interface and visualization tools can make predictions more understandable for researchers and healthcare professionals. Implementing mobile or cross-platform applications can enhance accessibility, allowing the system to be used in broader research, clinical, or experimental settings.

REFERENCES

[1] A. Hutke and J. Deshmukh, "Efficient model for early prediction of heart disease using ensemble technique," Proceedings of Engineering and Technology Innovation, 2025. [Online]. Available: https://doi.org/10.46604/peti.2024.14787

[2] R. Karmakar, U. Ghosh, A. Pal, S. Dey, D. Malik, and P. Sain, "A data balancing approach towards design of an expert system for heart disease prediction," arXiv preprint, 2024. [Online]. Available: https://arxiv.org/abs/2407.18606

[3] P. Muthulakshmi, M. Parveen, and P. Rajeswari, "Prediction of heart disease using ensemble learning," Indian Journal of Science and Technology, vol. 16, no. 20, 2023. [Online]. Available: https://doi.org/10.17485/IJST/v16i20.2279

[4] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," Scientific Reports, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-97547-6

[5] J. Tripathy, D. Sowjanya, C. Anilkumar, S.

Kunisetti, and Chandramouli, "Predictive modelling of Parkinson's disease progression using LightGBM classifier," Journal of Neonatal Surgery, 2025. [Online]. Available: https://www.jneonatalsurg.com/index.php/jns/article/view/2793

[6] I. Noushad and S. G. V., "Parkinson's disease prediction and prevention using machine learning," International Journal of Engineering Research & Technology (IJERT), 2024. [Online]. Available: https://www.ijert.org/parkinson-s-disease-prediction-and-prevention-using-machine-learning

[7] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, "Machine learning approaches to identify Parkinson's disease using voice signal features," Frontiers in Artificial Intelligence, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2023.1084001/full

[8] "Developing a model for Parkinson's disease detection using machine learning algorithms," Computers, Materials & Continua, 2024. [Online]. Available: https://doi.org/10.32604/cmc.2024.048967

[9] R. Mittal et al., "Machine learning approach to gait analysis for Parkinson's disease detection and severity classification," Frontiers in Robotics and AI, 2025. [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2025.1623529/full

[10] "Detection of Parkinson's disease using machine learning algorithms," Annual Methodological Archive Research Review, 2025. [Online]. Available: https://doi.org/10.63075/zzrnba26