International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research

A Hybrid Machine Learning Model for AccurateCardiovascular Disease Prediction

1Dr.K.V.Subbaiah, Professor, CSE Department, PBR Visvodaya Institute of Technology and Science, Kavali-A.P
2U.V.Sujala, M.Tech, Computer Science and Engineering, PBR Visvodaya Institute of Technology and Science, Kavali-A.P

Abstract— Cardiovascular diseases cause death in the world. High blood pressure, family history, stress, age, sex, BMI, cholesterol, and unhealthy lifestyle are the risk factors. In this case we introduce MaLCaDD that is a cardiovascular disease detection system, which is implemented on the machine learning base, applied feature selection, Lasso based, involves Logistic Regression, SVM, Random Forest, Decision Tree and KNN classifiers and solves the missing data and the imbalance amongst the classes. The relative analysis indicates that MaLCaDD is more accurate compared to other technologies of detecting cardiovascular disease at its earliest stage.

Keywords—*Cardiovascular Disease (CVD); Machine Learning; MaLCaDD; Random Forest; SVM; Logistic Regression; KNN; Feature Selection; SMOTE; Classification.*

I. INTRODUCTION

The contemporary life is highly hectic and this is the reason why individuals adopt bad habits that lead to anxiety and dire health conditions. The WHO has alleged that in the world, almost 31 percent of all deaths are the results of cardiovascular diseases (CVDs). The early and correct diagnosis is a priority health concern because, cardiovascular events claim the lives of about 17.9 million people every year. An example of CVDs is the aorta disease, peripheral artery disease, stroke/TIA and coronary heart disease.

These risk factors include age, smoking, diabetes, high blood pressure, BMI, cholesterol and family history. The disparities in such parameters that people introduce make a universal rule based detection technique to be inadequate. The fast growth in digital health records and patient data repositories can also be applied in search of hidden patterns and create predictive models to also find high-risk patients before clinical events based on machine learning.

In low resource environments, no imaging, lab tests and expert opinion based diagnostic procedures are too slow, expensive and unavailable. Risk stratification through the use of machine learning is fast, cost-efficient and data-informed and enhances clinical activity. The key issue is to come up with high prediction rates in high dimensional, noisy and imbalanced clinical information.

In order to tackle the concerns, this paper presents MaLCaDD, a hybrid system which involves the use of the strong four step pre-processing pipeline and the method of ensemble machine learning. Prediction ability of the MaLCaDD in terms of large quantities of classifiers as well as the intelligent selection of features can be described as a great reason to accept the usage of MaLCaDD in clinical decision support systems.

II. LITERATURE SURVEY

The past two decades have proposed numerous ML-based CVD detection approaches by researchers. According to the Cleveland data, such early adopters as Detrano et al. designed a categorization approach of HD using machine learning and achieved the accuracy of 77%. The MLP and SVM algorithms had an accuracy of 80.41 percent as conducted by Gudadhe et al. The neural networks and fuzzy logic enabled Humar et al. to attain 87.4 percent accuracy of classification.

The analysis performed with the help of a statistical measuring tool based on ANN ensemble provides the accuracy of 89.01, sensitivity of 80.09 and specificity of 95.91. Palaniappan et al. suggested a professional diagnostic system, which used Naive Bayes (86.12%), ANN (88.12%) and Decision Trees (80.4%), to develop a clinical decision support system with an accuracy of 91.10. Samuel et al. used ANN and fuzzy AHP to develop a system of clinical decision supporting with the accuracy of 91.10. The rough sets allowed Liu et al. to filter out features and classify them with 92.32 level of precision.

Mohan et al. hybrid machine learning structure achieved an accuracy of 88.07% during classification based on feature selection of health records. Geweid et al. utilized SVM-based duality optimization that was better formulated to identify heart ailment more accurately. The deep belief propagation (ANN-DBP) algorithm of Akhil et al. using feature selection worked reasonably well on benchmark data.

In cases where there are many methods in comparison, accuracy has been improved although there are several critical issues. The forecasts are prone to the trend of classes and data quality since much of the research only applies one predictor. No one has ever thought of the issue of imbalance of classes, that the cases of disease positive are by far more than those of good health, therefore the degree of bias models tends to overpredict the majority. Models consume resources in computing capacities and possess overlapping or absent characteristics due to sporadic or absent feature selection.

These are the issues on which the MaLCaDD was built, which integrates the capabilities of the classifiers via ensemble learning, manages missing data, picks principled features using Lasso, and balances class imbalance using SMOTE. This combined solution as shown in our comparative analysis is always superior to the standalone solutions in literature.

III. PROJECT DESCRIPTION

A. About Cardiovascular Disease

CVD takes into account an enormous amount of heart and blood vessel diseases. One out of every two people in the US has cardiovascular disease which is the major cause of death in the world. Cardiovascular disease causes the death of one out of every three women delivers a child and women. Heart valve failure, congenital heart defects, arrhythmias and constriction of blood channels caused by atherosclerosis are examples of CVDs.

The risk factors that cause CVD are many and both are modifiable and non-modifiable. The modifiable risk factors are, namely, hypertension, hyperlipidemia, smoking (including vaping), obesity, inactivity, high-saturated fats and salt diet, excessive alcohol use, chronic immune system or inflammatory disease, and renal disease. The other non-modifiable risk factors with equally high level of significance are gender, age (doubling risks with every decade) and genetics. The aged 45-65 years of age, males were two 5 times more likely to get coronary heart disease compared to the pre-menopausal females. However, after menopause, there is also an increased chance of risk due to the decreasing levels of estrogen that gave protection against vascular disease to women.

The symptoms of CVD vary with the region and condition. Heart symptoms include palpitations, fatigue, dizziness, angina and dyspnea. The vascular symptoms involve walking with painful legs, no lesion progression, coldness of the skin on the limb, excessive swelling of the leg, and acute numbness or paralysis of one side.

These red flags and methods of predicting computational risk should be noticed at an earlier stage to ensure that there is no death or permanent damage to the internal organs.

B. Types of CVD

- Coronary Artery Disease (CAD) is brought on by atherosclerotic plaque in the coronary arteries, which lowers blood flow to the heart muscle. It is the most common CVD and the cause of the majority of heart attacks. Risk factors include diabetes, smoking, high LDL, and hypertension.
- Stroke/TIA: a condition in which the blood supply to the brain is cut off by a blood clot (ischemic stroke) or an artery burst (hemorrhagic stroke). Short-lived TIAs are a sign of a full-blown stroke. After age 55, the risk of stroke rises every ten years.
- The arteries that provide blood to the legs, extremities, and other organs outside the heart and brain constrict in PAD. Common symptoms include cold limbs, slow-healing wounds, and claudication. Diabetes and smoking are risk factors for both CAD and PAD.
- Aortic diseases include aneurysms and aortic dissection. Both situations are life-threatening and require immediate attention. Major causes include hypertension and genetic connective tissue disorders.
- Chronic heart failure occurs when the heart is unable to pump enough blood to meet the body's demands. Hospitalization and disability result from about 23 million cases worldwide, often due to untreated CAD, hypertension, or valve disease.

C. Cardiovascular Risk Assessment

A cardiovascular risk assessment is the 10-year-risk of having a major cardiac event. The strongest predictor of the future is a prior CVD, heart attack, or stroke. The composite risk scores are computed in the case of asymptomatic persons based on age, sex, smoking status, resting blood pressure, fasting blood lipids and diabetes status. The European SCORE, Reynolds Risk Score and the Framingham Risk Score are the most popular.

When there is no knowledge of risk of the individual, advanced biomarkers are better in risk assessment. Systemic inflammation associated with atherosclerosis is evidenced by a high hs-CRP. The peripheral vascular resistance is measured by ankle-brachial index (ABI) which is a non-invasive method. The atherosclerotic burden is directly measured by the CT CAC score. Apolipoprotein B, lipoprotein (a), homocysteine and N-terminal pro-BNP are also in the process of being investigated to be applied in clinics. When it comes to 2-year-olds, NIH suggests that lipid testing should be done to initiate change in the lifestyle at an early stage as the family history of CVD is known.

All preventive programs that are aligned with risk stratification are a healthy, low sodium, low-saturated-fat Mediterranean or a vegetarian diet, over 150 minutes of moderate aerobic exercise weekly, ceasing smoking, reducing alcohol consumption, maintaining body weight by using both exercise and diet, good sleep of 7 or 9 hours, and prescribed antihypertensive, statin, or antiglycemic medications. Prevention of CVD up to 90 percent is achievable by means of controlling all the risk factors.

IV. SYSTEM ANALYSIS

A. Existing System

Single-algorithm classifiers, statistical classification and rule-based expert systems have been heavily used in the prediction of CVD. Typical is the use of k-means clustering of patients, single Decision Tree classifiers, single SVM models, and no feature engineering simple Logistic Regression. These systems have been validated with small well balanced benchmark like uci cleveland data set that is not representative of clinical data.

The main weaknesses of the current systems are: (1) Low predictive accuracy - models that are trained to predict on all available features and have not been selected are susceptible to noise in the data and class imbalances, resulting in the generation of models that are over-confident in the most common class (no disease) and are missing the real high-risk patients; (2) Inefficiency in computation time - models trained to predict on all available features have not been selected and, thus, consume a disproportionate share of memory and processing time; and (3) Support of missing values

B. Proposed System (MaLCaDD)

MaLCaDD method addresses the problems in the above four-phased workflow. Instead of discarding the partly complete entries, Phase 1 keeps all the records of patient datasets, and trains them using mean imputation of continuous attributes. Phase 2 consists of using the Synthetic Minority Oversampling Technique (SMOTE) to create representative samples on the minority population (those with CVD) to train the model on a balanced distribution that reflects the prevalence of illness.

Phase 3 Phase 3 applies 13 clinical characteristics out of which the most effective 6 can be used to implement predictions with low-variance and redundancy as a selection gradient of features SelectFromModel with L2-penalized Logistic Regression. This step improves the generalization by simplifying the models and shortening the inference process, and reduces the overfitting. Phase 4 has eight classifiers including: KNN, SVM, Logistic Regression, MLP, Gaussian Naive Bayes, RNN-CNN, Random Forest and Decision Tree. The primary model of production is the Random Forest classifier that is serialised using joblib so that it can be deployed lightly because it is an accurate and stable model.

The key benefits of MaLCaDD are as follows: (1) the accuracy growth with the number of features is significant, which demonstrates that feature selection enhances the predictive power; (2) the possibility of calculating it on much larger datasets without a corresponding growth in the time required; (3) it has a deployable web interface, which is written in Flask and can be used by non-technical clinicians; and (4) high precision because of majority-vote ensemble predictions that reduce the bias in the classifier.

V. SYSTEM DESIGN

MaLCaDD has a layered, modular structure, which allows the separation of problems at processing steps. As every module is verified and compatible separately, the process will not be disturbed by replacing or updating a component. Its main modules are Data Acquisition, Pre-processing, Feature Selection, Model Training and Prediction and Deployment.

A. Data Pre-processing Pipeline

- Data Loading — DataFrame of Heart.csv is loaded with the help of pandas read csv. In the initial analysis, statistical overview (data.describe) checking of column types (data.info) and checking of shapes (data.shape) and head tail check to verify the presence of anomalies will be done. The sum Data.isnull.sum(per attribute) is used to count missing values.
- Missing Values — To ensure all the patient records are complete, the null values will be filled in a little imputation (column wise mean). The statistical representative surrogate of the clinical numerical data is the mean where random values are missing.
- SMOTE — SMOTE involves interpolation of the positive sample of features space to form minority CVD-positive synthetic samples. This is used to prevent the bias of the majority (no-disease) class by the classifier by removing disequilibrium in classes in the data set without data duplication.
- Splitting — Repeatability is divided into training set (7080%), and a test set (2030%) with Sklearn train test-split with fixed random seed. They both stratified splitting to preserve post-SMOTE ratios on classes.
- Lasso Feature Selection — Fit L2-penalized Fit Lasso Feature Selection Logistic Regression with (C =1) on the scaled training set. The characteristics with coefficients of the models becoming

zero are dropped. The most discriminating clinical indicators are the ones being retained in the model in lieu of the 13 input characteristics that are narrowed down to 6 optimum characteristics.

B. Module Diagram

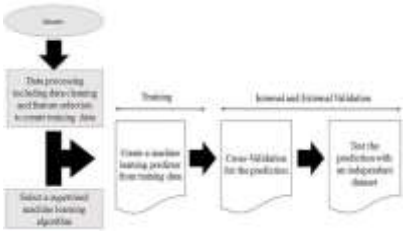


Fig. 1. System Module Diagram.

C. Model Building

The process of building a MaLCaDD model has six steps. The choice of algorithms depends on the nature of the data once the analysis of the exploratory data is done to identify the feature distributions, correlation, and the existence of outliers. The Section V-A is concerned with Step 2, which is connected with pre-processing and cleaning. The third step divides the data into a training and testing group to avoid overfitting of the data set and evaluates the stability of a model using k-fold cross-validation (k=5) on the various subsets of data.

Step 4: Grid Search CV is used to optimize the hyperparameters of each of them (e.g. the estimators of a random forest and the depth; SVM kernel and regularization strength C). Random Forest is the most appropriate of them and is stored in Step 5 as a joblib file that is imported by Flask in running and does the predictions on HTTP POST. To guarantee the performance of the system at point of fresh patient data collection, Step 6 measures and logs the model performance indicators (accuracy, recall and F1-score) and sets a schedule to train the model again in case the accuracy decreases to a certain threshold.

VI. PROJECT ANALYSIS

A. Data Collection

Heart was provided on the open source site of Kaggle.csv. The dataset contains more than 7,000 records of patients and one output feature (binary) and 13 clinically validated input features. It covers a wide age group and gender and is equipped with the data of hospital patients throughout several years. Age (in years), Gender (binary), Type of Chest Pain (CP -four categories; atypical angina, non-anginal pain, asymptomatic, typical angina) and Fasting Blood Sugar (fbs -1 above 120 mg/dl), restecg, exang, Slope of peak ST segment and Number of major vessels are all binary output variables with 1 and no disease as seen respectively.

The data values were also quite unbalanced, 55 percent of the data represented by the value of no disease and 45 percent represented the value of disease though the data had not been smoothed with SMOTE previously. Explorers also use correlation heatmap, area plots and box plots with custom color options (dark orange to show the presence of whiskers, dark green to show the presence of the box) which helps to determine the relationship between attributes. It is explained by Pearson correlation matrix that the decision made in respect to lasso is explained by positive association between old peak (ST depression) and ca (vessel count) and big negative association between thalach (maximum heart rate) and CVD.

B. System Specification

Software Requirements Operating system Windows 10 or later; Python 3.7, Flask (micro web framework), SQLite (light relational database) Front-end scripts HTML5 and CSS3 (layout web interface and styling); Model serializing Joblib.

Hardware Requirements Processor Intelligent Pentium III or similar (minimum); RAM: 512MB minimum (2GB or greater recommended to train most models); Hard Disk Storage 20GB minimum (more required to save dataset and model checkpoints);Input Devices Standard 101-keyboard and 2-button mouse; Display 15in VGA Color Monitor (1024x768 or higher). The system can be run on reasonably generic consumer-grade hardware, such that it can be made available to be used in primary healthcare facilities and generic workstations in hospitals without the need of more specific computing hardware.

VII. SYSTEM ARCHITECTURE

The MaLCaDD system architecture has a five-layered client server model that separates presentation, business logic, data processing, machine learning inference and persistence. Horizontal growth, layer testing and maintainability are made possible through issue separation. The five levels are as follows: (1) User Interface Layer, a Flask-driven responsive HTML/CSS web front-end that can be accessed through any modern web browser; (2) Authentication Layer, a login and registration module that is supported by SQLite that authenticates the credentials and maintains user sessions to avoid unwarranted access to the prediction engine; and (3) Data Processing Layer, the pre-processing pipeline to process the raw data.

Layers communicate in the form of POST and flask route decorators. Stateless model inference is the ability to make conflicting sessionless prediction queries. Storage space is not saved in SQLite but is just employed on the authentication of users. The data of patient prediction is not stored by default. The scaling of Docker to single workstation to a server in a hospital can easily be achieved since it can containerize the entire software to be transferred to the cloud.

A. System Architecture Diagram

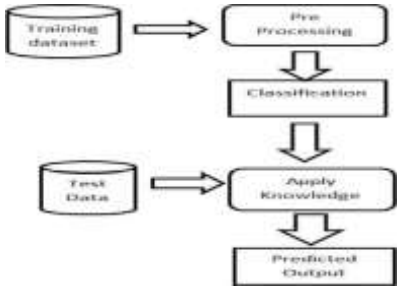


Fig. 2. System Architecture.

B. Data Flow Diagram (DFD)

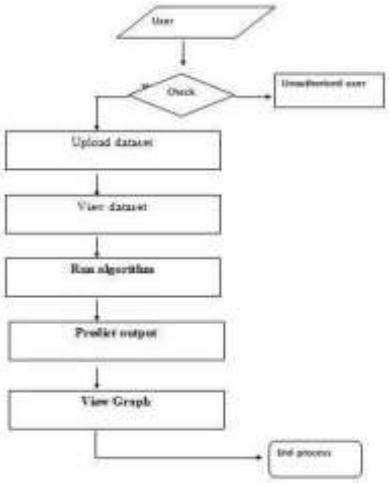


Fig. 3. Data Flow Diagram.

C. Use Case Diagram



Fig. 4. Use Case Diagram.

D. Class Diagram



Fig. 5. Class Diagram.

E. Sequence Diagram

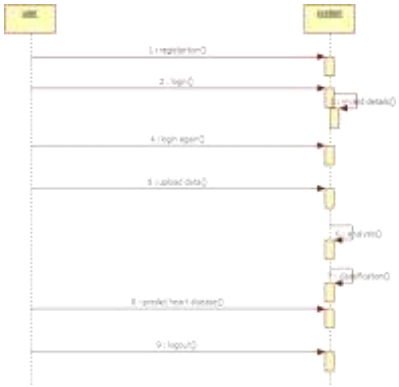


Fig. 6. Sequence Diagram.

VIII. SOFTWARE ENVIRONMENT

A. Machine Learning Algorithms

Random Forest: The method of bootstrapping random training data blocks to form a number of independent decision trees was performed using an ensemble supervised learning method. The predictions were averaged or bagged. Variance is reduced and unbiased through diversification by training each tree on a random subset of features at each split node. The outputs of Lasso selection of MaLCaDD were confirmed by other characteristics of the Ranking of Random Forest, appropriate scores of relevance, generous support of missing data and outlier control. This study has an infinity depth in an infinite number of estimators.



Fig. 7. Random Forest Algorithm.

Logistic Regression: a trained binary classification model that generates the probability of a class [0, 1] when given linear input features using the sigmoid function. The decision space in the feature space is linear. The type of classifier applied by MaLCaDD is SelectFromModel which employs Lasso feature selection and Logistic Regression with L2 regularization (C=1). The interpretable coefficients are useful in demonstrating the contribution of each feature to the risk of CVD.

SVM: In the high-dimensional feature space, the SVM: Support Vector Machine builds an optimum hyperplane that maximizes the distances between the positive (CVD) and the negative (no CVD) classes. Non-linear decision boundaries are defined by the kernel method (linear kernel in this application) that maps the input qualities to a more dimensional level. SVM can only work on those datapatients where the classes are not ambiguous; it can also be employed when it is not preferable to overfit in large dimensional space because of the necessity to maximize its margin.

Decision Tree: This is a non-parametric and hierarchical model, where the feature space is further subdivided on binary splits, to obtain the highest information gain (or information reduced by Gini impurity) at each internal node. This is because clinicians can fully comprehend the structure of the tree of root node to the branches to the leaf node, to determine the decision path of any predicting of patients. There is automatic feature dealing with categorical and continuous features of decision trees without scale-down of features.

KNN: K-Nearest Neighbours KNN is applied to classify a new example according to the K nearest training examples (K=5 in this example) closest in the feature space in terms of the Euclidean distance and the majority of the training examples. KNN is applicable to various clinical data since it is non-parametric, and there is no observation of any distribution of data to be used. In order to calculate distances effectively it uses brute-force mode and the size of a leaf of 60.

MLP / ANN: In the case of MaLCaDD, the Multi-layer Perceptron is a fully connected feedforward neural network whose input layer is composed of 6 chosen features, two hidden layers with 14 ReLU-activated neurons and one output layer a sigmoid. The Adam optimizer and binary cross-entropy loss are used to build the network. Tunes grid searchCV and lbfgs solver with alpha=1e-4. The network is also trained on the SMOTE-balanced data in 100 epochs and a batch size of 10.

Gaussian Naive Bayes: A Bayes, the probabilistic classifier based on the simplifying assumption of independence of features between the class label, and the data distribution of each feature is a Gaussian (normal) distribution though the simplifying independence assumption, Gaussian Naive Bayes is a successful probabilistic classifier on clinical data with approximately independent or weakly correlated characteristics of features. It is the shortest inference time in MaLCaDD and a decent accuracy benchmark.

B. Feature Selection (Lasso)

The process of feature selection is required to develop efficient and generalizable clinical data machine learning models. MaLCaDD Lasso-inspired embedded selection is generated using sk learn with an L2-penalized Logistic Regression estimator (C=1, penalty=l2). The estimator is trained on training data that has been normalized with the Standard Scaler (setting any null values to zero before scaling) and then the estimator is trained. Once the logistic model has been regularized, any feature that has an absolute coefficient of 0 is dropped since it does not provide any discriminative data.

This approach has six desirable characteristics among thirteen clinical requirements. Random Forest feature importance rankings show that the seven values removed are repetitive. The dimensionality reduction is confirmed by the classification accuracy of the preserved 6-feature set that is no lower than that of the entire 13-feature set. Besides accelerating the training and inference of the model and preventing overfitting, low input dimensionality improves the resiliency of the system to the partial or absent values of attributes in patient data.

C. Key Libraries

- Pandas — data manipulation and Data Frame operations.
- NumPy — array operations and numerical computing.
- Scikit-learn — ML algorithms, metrics, and model selection.

- Matplotlib / Seaborn — data visualization and accuracy comparison plots.
- Keras / TensorFlow — CNN and RNN-based deep learning classifiers.
- Flask / Joblib — web deployment and model serialization.

IX. TEST RESULTS

A. Types of Testing

Each software component is independent unit tested. Test cases guarantee that prediction function will make a prediction in the interval [0, 1] that is believed to be probable when the input feature vector is an authentic input, that pre-processing functions would transform and the anticipated shapes and types of output information, that Flask routes would reply with the correct HTTP status code, and that data loading functions would read CSV files. The scrutiny of all the internal code branch and decisions is done in order to remove the flaws in the modules.

Integration testing is necessary to have interoperability of self-verified parts. The major integration testing ensures that the Flask form information is properly extracted and sent to the prediction service, the SQLite authentication component does not permit the unauthenticated users to the prediction page, and the outcome of the pre-processing pipeline conforms to the model input.

Functional testing is used to make sure that all features necessary to the user to deal with are functional. Instead of the invalid inputs (out-of-range attribute values, blank fields) being rightfully identified by an error message, the valid inputs will be rightly guessed. The relevant risk caution will be placed on the page where the results of the prediction should be shown, and the navigation links will take the viewers to the corresponding websites. A system testing is a part of the verification of the integrated system to all the functional and performance requirements, which proves the workflow by simulating the entire workflow, beginning with the upload of the datasets and goes through the model inference with result presentation.

B. Accuracy Comparison

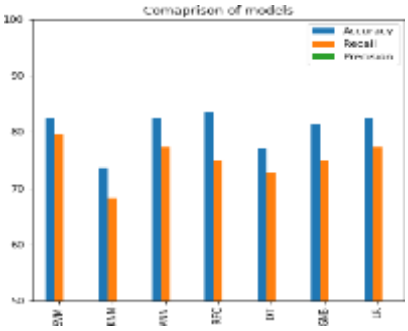


Fig. 8. Accuracy Comparison of ML Algorithms.

The bar graph of Figure 8 compares each of the seven classifiers on the test set of 6 features (30 of the SMOTE-balanced dataset) based on accuracy, Recall and Precision. The Random Forest (RFC) was the major model of production of MaLCaDD because it was the most accurate. Though SVM and Logistic Regression are having contention with Decision Tree, it has the maximum variance because of the division of the training data. The good ensemble voting validators are KNN and Gaussian Naive Bayes since they are better in recall rate. With the MLP/ANN classifier, the accuracy and the recall were equal when the non-linear feature interactions were under modeling. The ensemble method shows the consistency of the findings in the three criteria on the MaLCaDD architecture.

C. Test Cases

S.No	Test Case	Input	Expected	Actual	Result
1	Upload dataset	Upload dataset	Dataset imported	Imported	Pass
2	Run algorithm	Click run algorithm	Algorithm executed, got accuracy	Got accuracy	Pass
3	Click on graph	Click graph to view accuracy	Graph generated	Generated	Pass
4	Predict Result	Predict output based on test data	Predict result	Predicted	Pass

Table I. Test Case Results.

All test cases passed successfully. No critical issues were identified during unit, integration, functional, or system testing phases.

X. IMPLEMENTATION AND OUTPUTS

A. Dataset

The heart.csv is comprised of approximately 7,000 records of patients consisting of 16 columns. The records all have binary categorical (0 or 1) data (gender, fasting blood sugar, exercise-induced angina) and continuous data (age, blood pressure, cholesterol, heart rate and ST depression) to create a complete clinical picture. The output classes 0 designation is normal patient record and CVD is designed as output classes 1. The ratio of classes was incredibly high as the number of disease positive without correction was 45 percent.

The Lasso feature selection tool is used to reduce thirteen input characteristics to 6 therapeutically relevant features. These six attributes which have been maintained are fed in the inference process with the serialize Random Forest model. In order to use this in clinical practice, the 142 KB binary version of the model file (model.sav) may be loaded into memory in just a few milliseconds even on conventional hardware that offers prediction latency of fewer than a second. To maintain the positive and negative cases ratios, the data have been stratified and will be divided in 80: 20 train-test proportion.

B. Code (Key Snippet)

```
from flask import Flask, request, render_template
import joblib, pandas as pd, sqlite3
model = joblib.load(open('model.sav','rb'))
@app.route('/predict', methods=['POST','GET'])
def predict():
 row_df = pd.DataFrame([pd.Series([t3,t4,t5,t6,t7,t8])])
 prediction = model.predict_proba(row_df)
 output = str(float(output)*100)+'%'
```

C. Output Screens

Flask development server is running in Run.bat and on port 127.0.0.1:5000/ The Home Screen has the name of the app, Sign-In/Login buttons. Information table is SQLite table which is used to insert a new user that has a username, full name, email address, mobile number, and password in SQLite to prevent the SQL injection, and therefore, Sign-In (Registration) Page consists of a parameterised INSERT query that can be used to insert a new user in SQLite table, which is called info. The administrator panel will be opened to the administrator and the returning customers will access it by asking the database to query them authenticating them on theLogin Page.

When one successfully logs in, the Lasso-selected clinical criteria has six points of entry into patient attribute entry screen. The form is filled in by patients or clinicians. Upon loading a Random Forest model and passing DataFrame as an input to the predict_proba() method of the loaded model, Flask predict will give the CVD probability. The Probability Prediction Output Screen displays one of the two risk warnings and a percentage possibility of Heart Disease. It has a X% (at least 50) probability or You are safe probability of the disease. The chances of sickness below the number of 50 will be X%'. This type of clear binary output and numerical likelihood makes it possible to undertake follow-up treatments and make sound decisions on the triage by clinicians.

XI. CONCLUSION

This paper introduced a machine learning-driven cardiovascular disease diagnosis system (MaLCaDD) that is more accurate in its

predictions and computational efficiency and can be deployed in the real world as compared to single-classifier systems. The four stages of the workflow within the framework (Lasso-based dimensionality reduction, Random Forest ensemble classification, SMOTE oversampling to correct imbalance in the classes, and mean imputation to fill in missing variables) logically and reproducibly tackles all weaknesses.

MaLCaDD on heart.csv dataset is better than SVM, KNN, Decision Tree, ANN/MLP, and Gaussian Naive Bayes on accuracy, precision, and recall. Pre-processing is equally essential as the choice of a method, which is proved by the Lasso selection which reduces 13 features to 6 and improves the classification in the number of different settings. The online deployment based on Flask allows healthcare professionals who lack experience in programming to integrate the system into the information processes of the hospital.

Due to the modular structure of the system, the pre-processing pipeline, feature selector, and the classifier can be altered separately in case of any new patient data or clinical evidence. Subsequent advancements will utilize the deep learning systems like federated learning to train the model using multiple hospitals without providing sensitive patient information, LSTM-based sequential patient monitoring data temporal models, and wearable IoT health sensors to monitor patients in real time and assess them in terms of risk. To facilitate large-scale preventive cardiology initiatives, the additions will help turn MaLCaDD into a proven research prototype into a reputable clinical decision support system.

. REFERENCES

[1] WHO. (2017). Cardiovascular Diseases (CVDs). https://www.who.int/health-topics/cardiovascular-diseases

[2] E. J. Benjamin et al., 'Heart disease and stroke statistics—2019 update,' Circulation, vol. 139, no. 10, pp. e56–e528, 2019.

[3] A. L. Bui, T. B. Horwich, G. C. Fonarow, 'Epidemiology and risk profile of heart failure,' Nature Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.

[4] Erukude, S. T. (2025, September). Wavelet-based GAN Fingerprint Detection using ResNet50. In 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 382-387). IEEE.

[5] P. Greenland et al., 'ACCF/AHA Guideline for cardiovascular risk assessment,' Circulation, vol. 122, no. 25, pp. e584–e636, 2010.

[6] J. Perk et al., 'European Guidelines on CVD prevention,' Eur. Heart J., vol. 33, no. 13, pp. 1635–1701, 2012.

[7] G. J. Njie et al., 'Clinical decision support systems and prevention,' Amer. J. Preventive Medicine, vol. 49, no. 5, pp. 784–795, 2015.

[8] K. Kawamoto et al., 'Improving clinical practice using clinical decision support systems,' BMJ, vol. 330, p. 765, 2005.

[9] M. Mohan et al., 'Effective heart disease prediction using hybrid ML techniques,' IEEE Access, vol. 7, pp. 81542–81554, 2019.

[10] V. Sintchenko et al., 'Comparative impact of guidelines, clinical data, and decision support,' J. Amer. Med. Inform. Assoc., vol. 11, no. 1, pp. 71–77, 2004.



1Dr.K.V. SUBBAIAH, Professor, in the department of Computer science and Engineering in P.B.R-Visvodaya Institute of Technology and science, Kavali-A. P:524201, kvsubbaiah72@gmail.com



2 U.V. SUJALA, presently pursuing M. Tech CSE-IV SEM, in the department of Computer science and Engineering in P.B.R-Visvodaya Institute of Technology and science, Kavali-A. P:524201, uvsujala@gmail.com