Maximum Value Attribute-Based Extra Trees, XGBoost, and LightGBM for Early Disease Prediction

Babar Saeed[1], Basharat Ahmad Hassan[1], Khurram Gulzar[2], Noshi[3], Hisam Uddin[4]
[1]Department of Computer Science, Faculty of Computer Science & Information Technology, The University of Agriculture, Peshawar, 25130, Khyber Pakhtunkhwa, Pakistan.
[2]Universität Koblenz, Universitätsstraße 1, 56070 Metternich Germany
[3]School of Software Engineering, Dalian University of Technology, Dalian 116620, Liaoning Province, China.
[4]Department of National Institute of Peace and Conflict Studies, National University of Sciences and Technology (NUST), Sector H-12, Islamabad, 44000, Islamabad Capital Territory, Pakistan.
*Correspondence:basharat94@gmail.com

Citation | Saeed. B, Hassan. B. A, Gulzar. K, Noshi, Uddin. H, "Maximum Value Attribute-Based Extra Trees, XGBoost, and LightGBM for Early Disease Prediction", IJIST, Vol. 8 Issue.5 pp 2134-2162, August 2026
DOI | https://doi.org/10.33411/IJIST/1992
Received | Aug 05, 2026 **Revised |** Aug 17, 2026 **Accepted |** Aug 18, 2026 **Published |** Aug 22, 2026.

The evolution of intelligent healthcare systems has increased the importance of machine learning techniques in timely prediction of heart disease by using clinical and medical data. In recent years machine Learning techniques have been used widely in disease prediction systems due to their ability to examine complex medical patterns and also for better diagnostic accuracy. Many machine learning classifiers achieve strong predictive performance but still acquire high computational power and longer execution time during their training and testing phases. The ensemble learning algorithms such as Extra Trees, Extreme Gradient boosting and LightGBM have excellent performance in prediction of data due to their high classification accuracy and robustness. The main problem is that these models often suffer from high iteration counts, longer execution time and the inclusion of irrelevant or redundant features which may affect prediction efficiency and the overall performance of a model. To address these limitations, this research paper introduces a Rough Set Theory which is totally based on hybrid maximum value attribute (MVA). Hybrid MVA is a feature selection method that combines cardinality based ranking for categorical attributes and variance based ranking for continuous attributes in order to choose the most relevant features before training the models and predict heart disease in an efficient manner.The proposed models evaluate three techniques that are hybrid maximum value attribute Extra trees (MVA ET), hybrid maximum value Attribute XGBoost (MVA XGBoost) and hybrid maximum value attribute LightGBM (MVA LightGBM). This hybrid MVA method improves the performance of the model and improves computational efficiency by selecting the most relevant categorical and continuous features and removes irrelevant ones. The models are implemented using Python in Jupyter Notebook. The dataset is obtained from Kaggle platform which is titled as "Synthetic Heart Disease Prediction Dataset" that contains 50000 records and 20 features. The standard performance evaluation metrics including accuracy, precision, recall, F1 score and execution time are used to measure the success of the proposed approach. This report compares standard models against hybrid MVA across various train and test data distributions. The experimental evaluation is performed using accuracy, precision, recall, F1-score, execution time and iteration analysis to measure model effectiveness comprehensively and Experimental results show that the hybrid MVA based models outperforms the traditional classifiers in terms of prediction accuracy, computational efficiency and feature optimization. It reduces the feature dimensionality from 20 to 13 features, i.e., a 35% reduction in the input feature space before model training. For the Extra Trees classifier, the average iteration steps decreased from 46,787 to 31,040, corresponding to a 33.7% reduction in computational complexity. Similarly, training and execution time were also reduced in Hybrid MVA Extra Trees. The Hybrid MVA XGBoost model and Hybrid MVA LightGBM reduced the average iteration steps from 4,560 to 2,964, corresponding to a 35% reduction, while also reducing training and execution times. The Hybrid MVA Extra Trees model achieved an average accuracy of 99.07%, while Hybrid MVA XGBoost and Hybrid MVA LightGBM maintained average accuracies of 99.73% and 99.57%, respectively. Among the proposed approaches, Hybrid MVA XGBoost provided the best balance between predictive performance and computational efficiency.

Keywords: Disease Prediction, Machine Learning, Rough Set Theory, Maximum Value Attribute, Extra Trees, XGBoost, LightGBM, Feature Selection, Hybrid MVA, Clinical Symptoms

Introduction:

The increasing number of heart disease cases has placed a lot of stress on medical infrastructures. The rapid growth of patient data and the requirement for accurate clinical decision making have encouraged researchers to assist healthcare professionals in heart disease prediction and management. The timely diagnosis is particularly important as delayed detection of a disease can led to severe complications, high treatment costs and higher mortality rates.

In recent years the machine learning classifiers played an important role in predicting diseases including heart disease early and efficiently identify hidden patterns and its relationships within complicated medical datasets. However, there are several existing ML techniques that often require extensive hyperparameter tuning, multiple optimization stages and high iteration steps to achieve peak performance. These limitations increase computational cost and makes deployment very difficult in real time healthcare environments as in such cases quick response is very important.

Most modern ensemble and boosting algorithms have improved the prediction of diseases in medical diagnosis systems [1]. Among these approaches are Extreme Gradient Boosting (also known as XG Boost) is used widely for its good classification capability and robustness. It can handle high dimensional clinical data very efficiently [2]. XG Boost improves predictive learning from previous weak learners by correcting errors and makes it highly efficient by predicting complicated disease patterns and health of high-risk patient. Similarly the LightGBM has fast training speed, lower memory consumption and scalability for large medical datasets [3].

Despite these advantages there are many machine learning algorithms that still face several practical disadvantages in medical field [1]. There are many existing ML methods that needs extensive hyperparameter tuning, repeated optimization procedures and also high iteration steps to get best performance [1]. The computationally intensive procedures of these ML enhance execution time and resource utilization. It limits their applicability in real time diagnostic environments as rapid and timely prediction is very important in such critical cases. During feature selection and node optimization classical machine learning methods include complex mathematical calculations which further increases the procedural complexity [1]

To address these disadvantages, recent research has explored feature optimization and reduction techniques to make learning process easier and computational efficiency better. One solution is maximum value attribute (MVA) based on Rough Set Theory which provides an effective mathematical framework for managing uncertain, noisy and clinical data [1]. A recent study shows that RST-based feature selection utilizes the Quick educt algorithm which can reduce dimensionality while improving classification accuracy across different datasets such as heart disease and breast cancer [4]. RST has successfully reduced feature sets from 54 attributes down to just four while enhancing the performance of Support Vector Machines and Logistic Regression [4]. To enhance heart disease prediction, another study combined RST with Multi Criteria Decision Making which uses RST with Johnson's algorithm that separates the most important clinical features like age, blood pressure and serum cholesterol while also reducing model complexity and maintain predictive performance [5]. RST selects relevant features by reducing the number of features without affecting model performance [5].

The dataset of this research paper has categorical and continuous attributes that contain different data characteristics so if we use conventional MVA to apply a single ranking strategy to all attributes it will not produce the best feature subset. The categorical attributes contain a limited number of discrete values where conventional MVA can work

very well but continuous attributes have a wide range of numerical values where conventional MVA may not provide an appropriate ranking. Therefore, this study introduces a Hybrid MVA approach in which categorical attributes are ranked using the cardinality principle of rough set theory approach and continuous attributes are ranked using variance analysis. The selected features from both groups are then merged to form the final feature subset before model training. By choosing only the most informative attributes and removing redundant features, hybrid MVA based methods can reduce iteration steps, execution and training time while maintaining reliable predictive performance.

To overcome these challenges researchers are concentrating on lightweight and computationally efficient prediction frameworks that can produce high classification accuracy while reducing iteration steps. This study focuses on the problems of time consumption and complex procedural complexity in traditional ML models by integrating RST based MVA with extra trees, XG Boost and Light GBM. The proposed approach makes it easier to predict heart disease by selecting only the most relevant attributes instead of processing the entire feature space. By the reduction of features, we can improve both efficiency and interpretability in medical prediction systems. By just eliminating irrelevant and redundant information the proposed hybrid MVA method will minimize the computational cost, makes model execution faster and also improves overall model stability. In addition, it reduces feature dimensionality which can help healthcare practitioners like doctors and health researchers to better understand the contribution of important symptoms in prediction of heart disease. We integrate MVA with extra trees, XG boost and light GBM models that will helps to get a balanced tradeoff between predictive performance and computational efficiency [1].

The proposed framework is achieved using the heart disease prediction dataset based on symptoms and obtained from kaggle which contains symptoms-oriented data. The standard extra trees, XG Boost and Light GBM learning models are compared with hybrid MVA integrated using performance measures such as accuracy, precision, recall, F1 score, training time, execution time and iteration steps. The experimental findings shows that the proposed hybrid MVA based models have better predictive reliability while reducing computational complexity, training and execution time. These characteristics makes the proposed approach more suitable for intelligent healthcare systems and real time disease diagnostic applications where fast and accurate prediction is very important. The main objective of this research paper is the introduction of hybrid maximum value attribute feature selection method for separately processing categorical and continuous attributes according to their statistical characteristics because if we treat all features equally it will not show accurate result.

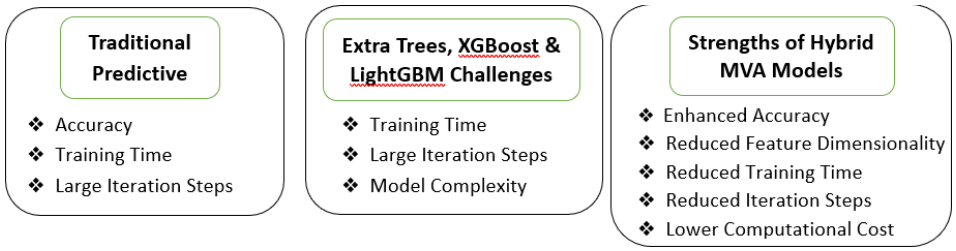

Figure 1. Issues in Standard Ensemble Models and Strengths of Proposed MVA Based Models

The Figure 1 shows the limitations of conventional predictive machine learning models and highlights the strengths of the proposed hybrid MVA by integrating with ensemble methods. The traditional predictive models may need significant training time and large iteration steps with low accuracy to get strong prediction performance. Although the ensemble machine learning techniques such as Extra Trees, XG Boost and Light GBM have

improved predictive capability for complex datasets but the problem is that these models still suffer from issues related to computational overhead, model complexity and increased number of iteration steps when processing high dimensional data. In healthcare prediction systems especially in heart disease prediction, reducing computational cost and maintaining high accuracy is very challenging task. The proposed hybrid MVA based Extra Trees, MVA XG Boost and MVA Light GBM models are introduced and integrated with Maximum Value Attribute (MVA) to reduce unnecessary computational steps, decrease model complexity and training time while improving prediction accuracy. Hybrid MVA processes categorical features which are ranked using the cardinality principle of RST and continuous features using variance-based analysis and selected features from both the categories are then combined to an optimized feature subset before model training. The proposed hybrid MVA integrated predictive model uses the optimized feature subset to reduce unnecessary computational steps provides a more efficient and reliable solution for heart disease prediction.

The main novelty of this research is Hybrid MVA feature selection technique based on RST for heart disease prediction. The conventional MVA methods process all attributes for feature selection strategy whereas the Hybrid MVA separately processes categorical and continuous features. Categorical attributes are ranked using the cardinality principle to measure their distinct values such as Gender, smoking, Alcohol use and diet and attributes with higher cardinality are considered for selection. The continuous attributes are ranked using variance analysis as it captures the variability present in numerical features like Age, Height, Diabetes or family history and features with higher variance are ranked as more informative. This hybrid selection strategy preserves relevant information from different data types while ignoring redundant attributes.

Another contribution of this study is the integration of the proposed Hybrid MVA with three ensemble learning algorithms and compare them with classical ensemble algorithms in terms of predictive performance and computational efficiency. The results shows that the proposed Hybrid MVA reduces the feature space while maintaining prediction performance and improving computational efficiency. These contributions make the proposed framework suitable for efficient heart disease prediction.

The remainder of this paper includes Section 2 which discusses the research background and related literature on extra trees, XG Boost and Light GBM learning models and disease prediction techniques. Then Section 3 presents the methodology of the proposed hybrid MVA integration with Extra Trees, XGBoost and LightGBM models. Section 4 presents the experimental setup, comparative analysis and performance evaluation results which are obtained from the heart disease prediction dataset. Finally, Section 5 concludes the research findings and discusses future research directions.

Related Work:

The use of AI in healthcare systems has undergone a significant paradigm shift over the last decade transitioning from traditional statistical modeling to advanced computational intelligence. Machine Learning has played an important role in medical science, particularly for prediction and classification of infectious and chronic diseases efficiently such as heart failure, Pneumonia, Dengue and Diabetes [1].

Evolution of Predictive Modeling in Epidemiology:

The disease forecasting depends heavily on traditional methods such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models [6]. While these models are very efficient for short term trend estimation they often have linearity and stationarity which restrict them to incorporate complex variables such as climatic conditions, demographic factors and medical diseases [6]. Consequently, researchers

have focused towards ML methods that can capture nonlinear relationships and execute high dimensional heterogeneous datasets with higher accuracy [6].

Logistic Regression:

Logistic Regression has been used widely to classify dengue risk categories in different provinces based on demographic characteristics of province and healthcare related indicators [6]. Due to its easy use and interpretability it is the most effective baseline model for patient disease risk and allowing health authorities to estimate disease probability and likelihood of severe or fatal dengue cases with acceptable accuracy [6]. A study was conducted on Indonesian provinces reported that the LR model has an accuracy of approximately 82% in predicting DENV infection and mortality outcomes [6]. However, despite being very efficient, LR has a disadvantage that when it handles high dimensional medical datasets and complex nonlinear relationships among variables. For this reason the researchers often consider it less suitable for advanced data analysis tasks and prefer more sophisticated machine or deep learning techniques to better represent the complexity of disease transmission and progression [6].

Support Vector Machines:

Support Vector Machines (SVM) have very strong performance in dengue diagnosis and epidemiological forecasting especially when integrated with advanced diagnostic technologies such as Raman spectroscopy to reagent free dengue prediction [7]. By analyzing unique molecular vibrations and biological signatures of the virus SVM get approximately 85% diagnostic accuracy and 93% specificity [7]. However the conventional SVM models are sensitive to class imbalance which means it often favors majority classes and reduces their efficiency in identifying severe dengue cases unless advanced preprocessing and weighting methods are applied to it [7]. In regional disease surveillance SVM has proven to be more effective for modeling complex nonlinear relationships in high dimensional datasets [8]. A study across Indonesian provinces reported that an SVM model with a Radial Basis function kernel has 83% accuracy which outperforms both Decision Tree and Logistic Regression models [8]. SVM has greater precision in identifying high risk regions making it very useful for patient disease risk and outbreak forecasting [8]

Moreover the Advanced variants such as Multiple Support Vector Machine Recursive Feature Elimination (MSVM RFE) have also been applied to identify molecular biomarkers linked to dengue severity [8]. Using this multivariate feature selection approach researchers has successfully identified cytokine profiles which makes it predict Dengue Hemorrhagic Fever with sensitivity and specificity above 80% [8]. The studies also shows that SVM based approaches can be weak with extremely imbalanced datasets sometimes producing biased predictions toward the dominant class when resampling techniques are not incorporated [8].

Decision Trees (DT)and Random Forests (RF):

The boosting machine learning methods including Decision Trees and Random Forest have outperformed traditional machine learning classifiers with respect to predict accuracy. These methods are recognized for their capability to minimize the common problem of overfitting by using bootstrap aggregation. Decision Trees have been used for their hierarchical structure that models complex, nonlinear relationships between clinical classifications and results. They are used for level prediction in Dengue and to detect necessary features in COVID19 infection. They are good for being simply interpretable by non-immunologist physicians. However, Standard DT algorithms like ID3 depends heavily on information Gain and entropy which are computationally intensive performance metrics for selecting the root node. Moreover, the standard DT models often involve complex measures that needs high iteration steps to reach best solution. DTs perform well on small categorical datasets but they face significant performance challenges when deal with larger,

high dimensional clinical datasets. The complex iteration cycles inherent in standard DT implementations are time consuming for real time diagnostic tools.

Random Forest is another ensemble machine learning classifier which has been used for population wide forecasting of Dengue burden at national scales. However, unlike Logistic Regression, RF is a nonparametric technique that cannot provides particular equations to examine relationships between counts and heterogeneous predictions [9]. Another severe limitation of RF regression is that it is not able to produce predictions that fall beyond the range of the training dataset for data points [9]. The RF model will inevitably under estimate cases in case of an unprecedented outbreak occurs [9]. It builds a forest of multiple trees and needs processing power and high iteration cycles as compared to simple models [9]. RF classifiers produce time lags when they were forecast to rapid changes in infectious disease dynamics [9]. The another difference from classical algorithms e.g., LR, SVR, Naïve Bayes and KNN) and Decision Trees, Random Forests identify relationships among attributes [10]. By executing these tree based models on large amount of datasets shows significant implementation and scalability challenges [10].

Extra Trees:

Recent research has highlighted the Extra Trees algorithm is a better learning method to predict heart disease which unlike traditional decision trees that looks for the optimal split at each node, extra trees selects splits randomly [11]. It is a technique that reduces overfitting and increases the model's robustness against noise in large datasets. This algorithm offers a critical balance between variance and bias to make it effective for processing high dimensional medical data.

In another study, the Extra Trees model using the UCI Machine Learning Repository shows a very good performance when integrated with Recursive Feature Elimination (RFE) to isolate the most important predictors such as BMI stroke history and general health. This integrated approach achieved an accuracy of 93.1%, a precision of 94.8% and a recall of 100% which consistently beats foundational benchmarks such as Logistic Regression (91.3%) SVM (91.5%) and Random Forest (91.3%). The Extra Trees architecture is increasingly viewed as a feasible component because of its flexibility and high predictive power for clinical decision support systems and allows healthcare providers to identify high risk cardiac patients with greater confidence and reliability [11].

Extreme Gradient Boosting (XGBoost):

Recent study which was conducted in Bangladesh has shown that Extreme Gradient Boosting as the premier machine learning classifier for forecasting dengue outbreaks among rapidly shifting infection dynamics [12]. In a comprehensive evaluation of multiple machine learning frameworks, XGBoost proved superior predictive reliability by achieving an extraordinary accuracy of 99.49%, a sensitivity of 99.24% and a specificity of 100% [12]. This exceptional performance is attributed to the algorithm's utilization of XGBoost where sequential weak learning algorithms like decision trees are implemented to correct the errors of previous iterations by minimizing the bias and capturing the complex non-linear relationships that were in clinical datasets. Furthermore, the study also highlights that if we integrate XGBoost with traditional time series models such as SARIMA it will give a more robust early warning technique than using individual architectures alone. These research paper study also shows that XGBoost is highly efficient to analyze high risk outbreak periods and offers a feasible and high precision solution for targeted medical interventions.

LightGBM:

Recent researches in medical diagnostics have highlighted Light Gradient Boosting Machine (Light GBM) as very effective classifier for high dimensional disease classification [13]. In a research study on thyroid disease detection, an improved architecture also called as Opti LightGBM was developed by combining sequential backward selection for selection of

features and whale optimization for feature weight refinement [13]. By using a leaf wise growth strategy to find out the highest splitter gain, the LightGBM lowers memory use and speedup computations relative to traditional level wise boosting techniques [13]. This Light GBM approach has gained a high accuracy of 99.75% which is better than conventional benchmarks such as Multi Kernel SVM and various neural network architectures. These studies highlight the potential of LightGBM in executing complex data and provides a robust balance between predictive precision and effective computations that are essential for real time diagnostic systems.

Maximum Value Attribute (MVA) and Feature Optimization:

MVA provides a mathematical method for data analysis and able to handle noisy and inaccurate data. The MVA approach relies on the rough value sets and the main purpose was to get higher cluster purity while minimizing computational cycles [14]. In this approach the attributes are totally depend on the cardinality of its value set. Furthermore, a rough set theory shows that attribute values can be used to explain and differentiate items within a dataset [14]. Therefore, the number of clusters formed by an attribute can be determined through the cardinality of its value set [14]. Compared to other existing techniques MVA has lower iteration steps and this improvement is mainly because the MVA technique uses the knowledge of the dataset including rough value sets, the number of clusters and rough indiscernibility relations to identify the most suitable clustering attribute [14].

The existing studies are summarized in Table 1 in terms of some relevant literature review parameters.

Table 1. Summary of Machine Learning Classifiers

Citation	Disease	Technique / Classifier Name	Accuracy / Metric	Advantages	Limitations
[11]	Heart Disease	Naive Bayes	84.3%	Probability approach often works well in practice	Makes the naive assumption that all features are conditionally independent
[11]	Heart Disease	KNN (K-Nearest Neighbor)	90.4%	Effective for finding common patterns in feature space	High dimensional data can lead to overfitting and needs significant memory for processing
[15]	COVID-19	ARMA (Auto Regressive Moving Average)	99.93%	Effective for short term trend estimation in small datasets	Insufficient for complex nonlinear forecasting &time-consuming during training
[16]	Dengue (Clinical Lab Data)	ANN & DA	90.91%	High performance even with small data i.e 77 records only	Small sample size may limit the generalizability of results
[17]	COVID-19	MLP (Multi-Layer Perceptron)	91.67%	Extract highly complex features are augmented by the integration of nonlinear activation functions	Often considered a black box model because it sacrifices interpretability for superior predictive power
[18]	Cerebral Infarction (Stroke)	CNN-MDRP (Convolutional Neural Network based Multimodal Disease Risk Prediction)	94.8%	Achieves faster convergence than unimodal models.	Performance drops if the number of text features is too small e.g<30.
[19]	Breast Cancer & Heart Disease	SVM	99.51% (Breast Cancer & 89.1% (heart disease)	Enables superior predictive performance high precision and robustness in image-based diagnostics	Large volume of data is required to train the model and also faces challenges in analyzing low contrast images
[20]	COVID-19	Random Forest (RF)	0.87	Highly effective for real time triage and identifying severe cases to optimize hospital resources	Single hospital system registry may limit generalizability did not include outpatient data

[21]	COVID-19	Decision Tree (DT)	94.99%	Achieved the highest accuracy among all tested models (LR, NB, SVM, ANN)	Sensitivity (89.2%) was lower than SVM and its specificity (93.22%) was lower than Naïve Bayes
[22]	COVID-19	Gradient-Boosting Machine (LightGBM)	0.90 (auROC)	Predicts COVID-19 using only eight simple binary features obtained via basic questions	Relies on self-reported symptoms which are subject to bias
[23]	Oral Cancer	Extra Tree	98%	Introduces randomness in feature/threshold selection to reduce overfitting and improve generalization	Requires validation in larger more diverse cohorts to ensure generalizability
[24]	COVID-19	MRPMC (Ensemble Model)	92.4%	Integrates LR, SVM, GBDT and NN via weighted voting for maximum precision	Model stability depends on the specific weights assigned to the sub-models (LR: 0.25, SVM: 0.3, GBDT: 0.1, NN: 0.35)
[25]	COVID-19	HPO-FS-SVM (SVM optimized with MRMR and Modified Cuckoo Search)	96.73%	The MRMR algorithm identifies the most significant traits while reducing feature redundancy	Heavily dependent on the specific type of feature selection and hyper-parameter optimization chosen
[26]	Chronic Kidney Disease (CKD)	CatBoost	98.75%	Best overall performance and showed highest accuracy compared to other models in research	Success is tied to specific nature inspired preprocessing steps requires more validation on diverse datasets
[27]	Diabetes	Stacking Ensemble (XGBoost + KNN with LightGBM meta-learner)	94.82%	demonstrates high stability on new data with minimal training-testing gaps	oversampling techniques (SMOTE + Tomek Links) may not perfectly reflect real world distributions
[28]	COVID-19	AdaBoost	92%	Effective for categorizing patients into supervision or monitoring levels	Accuracy may be affected by smaller datasets and reliability increases with higher patient sample counts.
[29]	Breast Cancer	Hoeffding Tree	97.9%	Capable of learning from massive data streams with lower memory consumption	Spends extensive computational time inspecting for ties in the algorithm

[30]	COVID-19	Prophet Algorithm	R^2: up to 1.00 (data perfectly fits the prediction line)	Enables fast feasible real-time outcomes via cloud-based mining	Predictive capability can fluctuate depending on the specific sliding window (round) analyzed
[31]	Coronary Artery Disease	Bayesian-Optimized SVM	97.67%	Achieves zero false negatives systematic and efficient hyperparameter search compared to grid or random methods	Heavily integrated with a hybrid feature selection approach using AdaBoost and Decision Trees to identify relevant predictors
[32]	Diabetes, CVD and Cancer	CNN-LSTM (hybrid)	95.3%	Captures both spatial features (via CNN) from images and temporal sequences (via LSTM) from clinical records	High computational requirements and complex architecture need massive amounts of high-quality labeled data.
[33]	Alzheimer's Disease	Quadratic Discriminant Analysis (QDA)	99.9%	Provided optimal diagnosis performance with near perfect accuracy	Requires further validation with a larger more diverse public dataset to ensure robustness.
[34]	Heart Disease	J48 Decision Tree	97.73%	Highly efficient in detecting benign and malignant risk	Its accuracy is sensitive to the pruning process and selection of attributes like age and blood pressure.
[35]	Parkinson's Disease (Biomedical voice measurements)	Gradient Bosting Decision Tree (GBDT)	98.62% (ROC AUC)	provides a more transparent decision-making process than black box DL models	Computational complexity and training time can increase significantly in very high-dimensional feature spaces
[35]	Thyroid Abnormalities	Elman Neural Network (ENN)	97.8%	Making it ideal for modeling nonlinear dynamic systems like thyroid function	requires additional layers and weight distributions to manage previous state information
[36]	Cardiotocography (CTG)	Hybrid Cascade Forward NN (HECFNN)	99.25%	Fuses the benefits of both Cascade Forward (CFNN) and Elman (ENN) networks	Model success is highly sensitive to the number of hidden layers and neurons

Recent Feature Selection and Explainable AI Studies:

Recent studies are largely focusing on combining feature selection with machine learning models in order to improve predictive performance and lower computational cost [37]. A recent study on Mitral Regurgitation (MR) shows a semi-automatic pipeline extract four dimensional morphological features from cardiac cine MRI scans [37]. A total of 187 candidate features were extracted from the reconstructed mitral annulus and implemented the minimum redundancy maximum relevance (mRMR) technique in order to identify the 25 most relevant clinical markers [37]. These selected features were evaluated using linear discriminant analysis LDA and random forest RF with accuracies of 72% and 73%, respectively [37]. In another research, the hybrid feature selection approach for early detection of chronic liver disease has merged with Pearson correlation coefficient, Gain Ratio and Relief with projection techniques that includes principal component analysis and linear discriminant analysis [38]. Moreover, SHAP technique identifies features that contributed more strongly to diagnostic predictions and the proposed approach by using a four hidden layer deep neural network has achieved an accuracy of 90.12% and an average accuracy improvement of 13.93% over conventional predictors has been reported [38].

The CNN-LSTM, GB-NN, and AE-SVM architectures combine with CNN LSTM has achieved the highest reported accuracy of 95.3% [39]. The integration of SHAP gives additional interpretability by extracting important clinical variables such as age, BMI and blood pressure [39]. This combination of predictive modeling and Explainable AI actually shows the importance of diagnostic healthcare systems to understand the factors contributing to model predictions [39]. Another research has used the Bom Gene algorithm that implements a two staged hybrid feature selection method which combines Boruta and mRMR to optimize high dimensional gene expression datasets [40]. This technique has achieved maximum accuracy of 98.113% while reducing training time and ensuring high stability across 25 diverse biological datasets [40].

Methodology:

This study uses the heart disease prediction dataset which was taken from Kaggle. The dataset has 50,000 patient records including 20 clinical features. The attributes in dataset include variables such as age, gender, smoking, alcohol intake, physical Activity, diet, stress level, hypertension, diabetes, hyperlipidemia, family history, previous heart attack, systolic blood pressure, diastolic blood pressure, heart rate, blood sugar fasting, cholesterol total and all these symptoms are used as attributes for prediction of a heart disease. The target variable is heart disease with a binary class where 0 shows the absence and 1 represents the presence of a heart disease.

The dataset in this research study contains a reasonably balance class distribution with 26,827 samples belonging to the non-heart disease class and 23,173 samples for heart disease class that makes it more suitable for reducing class bias and also supports reliable model training and evaluation. The dataset has been examined before model training to find out missing and inconsistent entries and duplicate records and most of the attributes contains complete records was found in the dataset but the alcohol intake attribute contained 20,109 missing values. The missing values of the corresponding feature was removed before model training. After that the data preprocessing was performed and Label Encoding was applied to convert categorical class labels into numerical form. The feature selection was carried out for proposed hybrid MVA models to extract the most relevant and cardinality based top ranked features for categorical features whereas variance ranking for continuous features in order to reduce dimensionality and improve computational efficiency. After data preprocessing the dataset was divided into 90/10, 80/20 and 70/30 training and testing sets. All the models were trained using 76 estimators with a fixed random seed of 42 to ensure reproducibility of the experimental results. The standard Extra Trees, standard

XGBoost, standard LightGBM, hybrid MVA Extra Trees, hybrid MVA XGBoost and hybrid MVA LightGBM models were implemented in python using Jupyter Notebook. Different libraries, including Pandas, NumPy, scikitlearn, XGBoost and LightGBM, have been used for data preprocessing, visualization, feature selection, model implementation and heart disease prediction.

Standard Machine Learning Models:

The following high-performance algorithms serve as benchmarks:

Standard Extra Trees: The Extra Trees classifier is implemented as one of the baseline models in our study for comparison with the hybrid MVA approach which builds multiple randomized decision trees and merge their predictions to improve classification performance. It randomly decides split threshold to improve model diversity while manages overfitting. The quality of each split is evaluated using the Gini impurity criterion, which is defined as

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2 \quad (1)$$

where $p_i$represents the probability of class $i$and $c$is the total number of classes. The average depth and number of features of all generated decision trees was calculated to determine the computational complexity of extra trees by using

Iterations=Number of Trees × Number of Features ×Average Tree Depth (1)

Where, 76 trees were constructed by utilizing 20 input features and the average tree depth was evaluated from all trained trees.

XGBoost: Extreme Gradient Boosting also known as XG Boost is another machine learning classifier that builds decision trees sequentially to correct errors. XG Boost is used widely for medical diagnosis and heart disease prediction. XGBoost regularization techniques reduces model complexity and improve generalization performance. Like the extra trees it is also implemented as the baseline model in our study for comparison with the hybrid MVA approach.

The main function of XGBoost is a loss function and a regularization term which can be expressed as follows:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \quad (2)$$

where

$l(y_i, \hat{y}_i)$is the prediction loss,

$f_k$represents the k^{th}decision tree,

$\Omega(f_k)$is the regularization term used to control model complexity.

The regularization function is defined as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (3)$$

where

$T$is the total number of leaf nodes,

$w_j$is the weight of the j^{th}leaf,

γcontrols tree complexity,

$\lambda$is the L2 regularization parameter.

During tree construction, XGBoost selects the best split by maximizing the Gain function,

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \quad (4)$$

where

G_L, G_R are the first-order gradients,
H_L, H_R are the second-order Hessians,
λ and γ are regularization parameters.

To estimate the computational complexity of the model, the average depth of all generated trees is calculated and the approximate iteration steps are computed as

$$Iteration = N_{estimators} \times N_{features} \times AverageDepth$$

Where
$N_{estimators} = 76$,
$N_{features} = 20$,
Average Depth is obtained from all trained boosting trees.

The implementation of XGBoost uses 76 boosting trees with a learning rate of 0.05 and a maximum tree depth of 3. The complete set of all 20 input features are used for model training.

LightGBM: Light Gradient Boosting Machine (LightGBM) is a ML framework that focused mainly on speed and efficiency in large feature spaces. It is a gradient boosting framework that is based on decision trees which builds trees using a leaf-wise growth strategy and allows the model to achieve higher prediction accuracy while requiring fewer iteration steps. LightGBM is particularly suitable for high dimensional datasets as it reduces training time and manage predictive performance. During tree construction, LightGBM selects the best split by maximizing the information gain, which is computed as

$$Gain = \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma \qquad (5)$$

where
G_L and G_R denote the first-order gradients of the left and right child nodes,
H_L and H_R represent the corresponding second-order Hessians,
λ is the L2 regularization parameter,
γ controls tree complexity by penalizing unnecessary splits.

The computational complexity of the proposed LightGBM implementation is estimated using

Iterations=Number of Trees × Number of Features × Average Tree Depth (6)

where the total 76 boosting trees were constructed using 20 input features and the average tree depth was evaluated from all trained trees by LightGBM model. The model was implemented with a learning rate of 0.05 and a maximum tree depth of 3.

Hybrid MVA Feature Selection:

The hybrid maximum value attribute technique combines categorical and continuous attributes before model training. For categorical attributes, the MVA based on Rough Set Theory is applied to select top 2 features out of 6 in the dataset according to their cardinality values. For continuous attributes, a variance based ranking approach is applied to select 11 features with higher variability. These continuous attributes give more discriminative information for classification. The selected categorical and continuous attributes are finally merged to make the final feature subset. It ignores irrelevant data thereby decreasing the iteration steps, training and execution time.

Standard vs MVA Based Models:

In this study two categories of predictive models were developed and compared which are standard machine learning models and MVA based optimized models. The standard models were trained using all 20 input features in the dataset allowing the classifiers to learn from the complete feature space but the hybrid MVA based models uses a two stage feature selection strategy i.e categorical and continuous attributes and selected features from both of these are then combined to form the optimized subset of features used for model

training. For every categorical attribute, attributes having larger cardinality values are ranked higher.

$$Cardinality(A_i) = | V(A_i) | \quad (7)$$

Where

$A_i$denotes the categorical attribute.

$V(A_i)$represents the set of unique values.

$| V(A_i) |$is the cardinality.

For each continuous attribute, features with higher variance preserve greater variability within the dataset.

$$Var(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \quad (8)$$

Where

$x_i$is an individual observation,

μdenotes the mean,

$n$represents the total number of observations.

Finally, the merged feature subset is then used for training the Hybrid MVA models.

$$F_{Hybrid} = F_{Cardinality} \cup F_{Variance} \quad 9$$

Where

$F_{Cardinality}$is the subset selected using MVA.

$F_{Variance}$is the subset selected using variance ranking.

The purpose of integrating hybrid MVA was to reduce dataset dimensionality and ignores less relevant attributes while preserving important diagnostic information required for heart disease prediction. This feature optimization approach was expected to reduce training time, execution time and iteration steps. Furthermore, the study also identifies whether by reducing the features in hybrid MVA models can maintain predictive performance comparable to the standard models in terms of accuracy, precision, recall and F1 score. In addition, the training time, execution

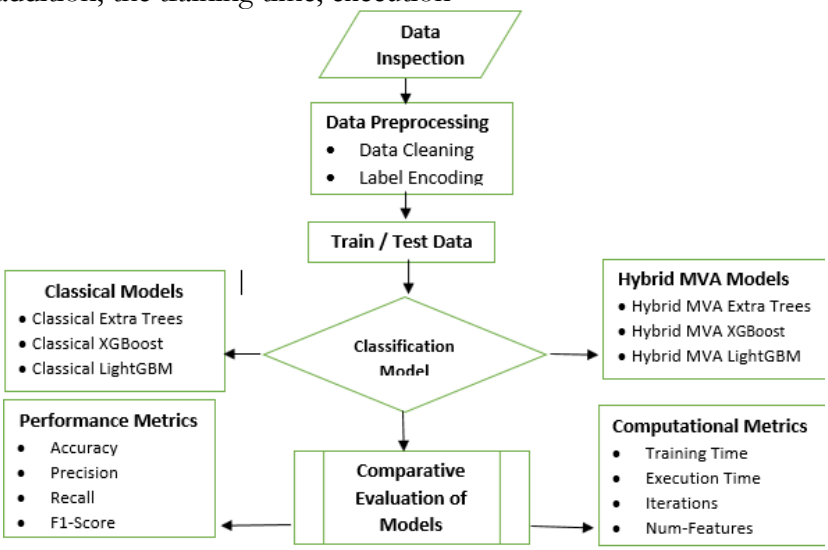

Figure 2. Proposed Methodology Flow chart

The proposed methodology of integrating hybrid MVA techniques with machine learning models is illustrated in the flowchart presented in Figure 2. The process includes different phases such as data inspection, data preprocessing and setting train/test data for model development. During preprocessing, data cleaning and data selection are performed to improve data quality and reliability. After this the Hybrid MVA feature selection process divides the features into categorical attribute where the cardinality of each attribute is

calculated based on the number of distinct values and in case of continuous features variance is calculated to measure the amount of variation within each attribute. The attributes are divided into categorical and continuous attributes because they have different data characteristics so they also need a different ranking criterion. After that they both are combined to form the final Hybrid MVA feature subset and then the dataset is divided into training and testing sets. The methodology implements both standard models (Extra Trees, XGBoost and LightGBM) and MVA based models (hybrid MVA Extra Trees, hybrid MVA XGBoost and hybrid MVA LightGBM) to perform classification on heart disease based on performance metrics and computational metrics.

Experimental Setup:

The research process is systematically divided into data collection, preprocessing, training and a comparative performance analysis of standard and hybrid MVA integrated models.

Implementation Environment:

The experiments were performed on high performance system for consistent evaluation of efficiency metrics. Following are the details of system that have been used for the prediction of different heart diseases.

Hardware Specifications: HP 11th Gen Intel(R) Core (TM) i51145G7 @ 2.60GHz 2.61 GHz with 16GB RAM.

Software Environment: Jupyter Notebook (version 7.4.5) was utilized as the primary implementation toolkit.

Libraries: The predictive models were developed using the Python repository calling upon Pandas for data visualization and scikit-learn, XGBoostand LightGBM libraries for algorithm implementation

Dataset Description and Clinical Attributes:

This research paper uses the dataset of heart disease prediction dataset that has been downloaded from Kaggle open-source website.

Dataset Shape: The raw information consists of 50000 patient records with 20 input features representing medical history and laboratory measurements.

Attributes: The predictors include input features to predict heart disease such as age, gender, weight, height, bmi, smoking, alcohol intake, physical activity, diet, stress level, hypertension, diabetes, hyperlipidemia, family history, previous heart attack, systolic blood pressure, diastolic blood pressure, heart rate, blood sugar fasting, and cholesterol total.

Target Labels: The target attribute is heart disease where 0 indicates the absence of heart disease and 1 indicates the presence of heart disease.

Data Cleaning: Null values and duplicate records were examined initially to ensure a verified and precise dataset for model training and missing values were identified in the Alcohol Intake attribute. Moreover, preprocessing and label encoding were also implemented before model training.

Data Distribution, Splitting and hyperparameter settings:

The dataset has been categorized into three training and testing splits to evaluate the reliability and efficiency of the models across different data. One is 90/10 split, which means that 90% data will be for selected for training of model and 10% for testing. The second is 80/20 split which means that 80% data is for training and 20% for testing. The last one is 70/30 split which means that 70% data has been selected for training and 30% for testing.

Table 2. Hyperparameter Configuration of Standard and Hybrid MVA Models

Hyperparameter	Extra Trees / Hybrid MVA ET	XGBoost / Hybrid MVA XGBoost	LightGBM / Hybrid MVA LightGBM
Number of estimators	76	76	76
Criterion	Gini	Gradient based gain	Gradient based gain

Objective	_	Binary logistic	Binary
Maximum depth	None	3	3
Minimum samples split	2	—	—
Minimum samples leaf	1	—	—
Learning rate	—	0.05	0.05
Number of leaves	—	—	31
Minimum child samples	—	—	20
Early stopping	Not used	Not used	Not used

Table 2 shows the complete hyperparameter settings for the three classification algorithms. The Extra Trees classifier used the Gini criterion, 76 estimators and a minimum of samples for splitting is 2 and 1 sample has been used per leaf. The XGBoost used 76 estimators with learning rate of 0.05, maximum depth of 3 and for evaluation metric it uses logloss. In case of LightGBM 76 estimators were used with a learning rate of 0.05 and maximum depth of 3 whereas the number of leaves used were 31 and 20 minimum child samples. No early stopping was applied; therefore, training was stopped after the predefined 76 estimators. The same configurations were used for both standard and Hybrid MVA ensemble models.

Performance Classification Criteria:

The performance of models has been checked and ranked using the following standard performance metrics and computational parameters. The following are mentioned below to measure predictive accuracy, efficiency and optimization capability:

Accuracy (%): The amount of total number of correctly classified diseases by the model out of the total cases examined is called accuracy.

Precision (%): The precision means that the model evaluates the correctness of positive predictions in data.

Recall (%): Recall is actually the ability of a model to find out all actual positive cases in data.

F1 Score (%): The average of precision and recall that gives a balanced reliability measure.

Training Time: The training time is the total time required to train the model before making predictions.

Execution Time: The total time taken in minutes/seconds to create the prediction result.

Iteration Steps: Iteration steps are very important and refer to the number of mathematical computational cycles required for the mode.

Average Depth: The mean depth of all decision trees generated during model training which is used to calculate the computational complexity of models.

Number of Features: Number of features shows the total number of features that are selected by the model during the classification process.

Experimental Results and Discussion:

Dataset Sample Representation:

Table 3(a). First five rows of the dataset

Age	Gender	Weight	Height	BMI	Smoking	Alcohol	Activity	Diet	Stress
48	Male	78	157	26.4	Never	NaN	Sedentary	Healthy	Medium
35	Female	73	163	33.0	Never	Low	Active	Average	High
79	Female	88	152	32.3	Never	NaN	Moderate	Average	Medium
75	Male	106	171	37.4	Never	Moderate	Moderate	Average	Low
34	Female	65	191	18.5	Current	NaN	Sedentary	Healthy	Low

Table 3(b). Remaining first five rows 1

Diabetes	Hyperlip.	Family	Prev HA	SysBP	DiaBP	HR	Sugar	Chol	Target
0	1	1	0	104	99	71	165	200	0

0	1	1	0	111	72	60	145	206	0
0	0	1	0	116	102	78	148	208	0
0	1	0	0	171	92	109	105	290	1
1	0	0	0	164	67	108	116	220	1

Table 3(a) and Table 3(b) shows the first five rows of the dataset that was downloaded from Kaggle website named heart disease prediction dataset. Each row represents the medical record of an individual patient which includes the patient lifestyle factors, medical history, vital signs and laboratory results. The dataset consists of 50000 patient records with 20 input features and one output variable (heart disease) where 0 represents the absence and 1 indicates the presence of heart disease. There are 26827 records without heart disease i.e Class 0 and 23173 records with heart disease i.e Class 1. This data structure is used for training and evaluating the prediction of patient diseases using proposed models.

Hybrid MVA Feature Ranking for Categorical and Continuous Attributes:

Table 4. Hybrid MVA Attributes

Categorical	Card.	Continuous	Variance
Gender	2	Age	208.4584
Smoking	3	Weight	408.5758
Alcohol_Intake	4	Height	207.9473
Physical_Activity	3	BMI	40.5450
Diet	3	Hypertension	0.2099
Stress_Level	3	Diabetes	0.1596
		Hyperlipidemia	0.1883
		Family_History	0.2401
		Previous_Heart_Attack	0.0894
		Systolic_BP	532.85
		Diastolic_BP	297.8407
		Heart_Rate	209.9985
		Blood_Sugar_Fasting	1004.3516
		Cholesterol_Total	1862.5669

Table 4 presents the categorical and continuous attributes in the hybrid MVA technique in which two separate ranking strategies were employed. The categorical attributes were ranked using the MVA cardinality principle based on RST and continuous attributes by their variance rankings. The top 2 categorical attributes and 11 continuous attributes were selected and then merged to form the final feature subset which decreases the dimensionality of the dataset, lowers computational complexity and training time while giving high predictive performance.

Table 5. Ablation Analysis of the Hybrid MVA Framework

Method	90/10 Accuracy	80/20 Accuracy	70/30 Accuracy	Interpretation
Cardinality-only	0.7228*	0.7165*	0.7219*	Uses only categorical/cardinality-based informatierroron and gives relatively weak performance
Variance-only	0.7238*	0.7187*	0.7221*	Uses only variance information from continuous attributes and also gives weak performance
Hybrid MVA	0.9814	0.9849	0.9865	Combines cardinality-based and variance-based information and substantially improves predictive

				performance

According to table 5, the ablation study demonstrates that neither cardinality based ranking approach nor variance based alone are giving good predictive performance. Both approaches strategies are giving the accuracies of approximately 0.72 across the three train test splits. However, in contrast, the proposed Hybrid MVA framework achieved accuracies of 0.9814, 0.9849, and 0.9865 for the 90/10, 80/20, and 70/30 splits respectively. This improvement is the proof that the combination of two ranking methods is giving strong predictive performance while also lowering computational efficiency.

Table 6. Categorical feature selection

Categorical Features	Continuous Features	Total Features	Accuracy	Iterations
2	11	13	0.9913	30862
3	11	14	0.9875	34230
4	11	15	0.9849	35505
5	11	16	0.9920	36880
6	11	17	0.9900	39270

Table 6 shows that the number of categorical features effects on computational complexity and slightly on classification performance. By using 80/20 train and test split, the selection 2 categorical and 11 continuous features achieved an accuracy of 0.9913 with 30862 iteration steps. The Hybrid MVA Extra Trees model was used for this analysis shown in table 4 while additional experiments with Hybrid MVA XGBoost and Hybrid MVA LightGBM also achieved approximately 99% accuracy while reducing iteration steps. Therefore, we select 2 categorical attributes for continuous analysis with number of features from 7 to 11 as shown in table 7 and it showed improved accuracy from 0.72 to 0.99 while increasing the number of iterations from 22,275 to 30,862. Therefore, we selected 2 categorical and 11 continuous features due to its strong predictive performance and computational complexity which achieves a 99.13% accuracy while also reduced number of features to just 13.

Table 7. Categorical feature selection

Categorical Features	Continuous Features	Total Features	Average Accuracy	Average Iterations
2	7	09	0.720	22275
2	8	10	0.868	24730
2	9	11	0.853	27467
2	10	12	0.949	29364
2	11	13	0.9913	30862

Table 7. Results of Standard Extra Trees

Split	Accuracy	Precision	Recall	F1-Score	Train Time (s)	Execution Time (s)	Iterations	Avg. Depth	No. of Features
90/10	0.9864	0.9864	0.9864	0.9864	6.5177	6.7728	47,560	31.29	20
80/20	0.9853	0.9853	0.9853	0.9853	5.7963	6.1972	46,460	30.57	20
70/30	0.9859	0.9859	0.9859	0.9859	5.0495	5.5785	46,340	30.49	20

Table 8. Results of Hybrid MVA Based Extra Trees

Split	Accuracy	Precision	Recall	F1-Score	Train Time (s)	Execution Time (s)	Iterations	Avg. Depth	No. of Features
90/10	0.9910	0.9910	0.9910	0.9910	4.8970	5.1267	31,655	32.04	13
80/20	0.9913	0.9913	0.9913	0.9913	4.3341	4.6483	30,862	31.24	13
70/30	0.9899	0.9899	0.9899	0.9899	4.2478	4.7500	30,602	30.97	13

Table 10. Results of Standard XG Boost

Split	Accuracy	Precision	Recall	F1-Score	Train Time (s)	Execution Time (s)	Iterations	Avg. Depth	No. of Features
90/10	0.9962	0.9962	0.9962	0.9962	0.3331	0.4433	4,560	3.00	20
80/20	0.9956	0.9956	0.9956	0.9956	0.2981	0.4423	4,560	3.00	20
70/30	1.0000	1.0000	1.0000	1.0000	0.2709	0.4126	4,560	3.00	20

Table 11. Results of Hybrid MVA Based XG Boost

Split	Accuracy	Precision	Recall	F1-Score	Train Time (s)	Execution Time (s)	Iterations	Avg. Depth	No. of Features
90/10	0.9962	0.9962	0.9962	0.9962	0.2685	0.3651	2,964	3.00	13
80/20	0.9956	0.9956	0.9956	0.9956	0.2506	0.3825	2,964	3.00	13
70/30	1.0000	1.0000	1.0000	1.0000	0.2412	0.3841	2,964	3.00	13

Table 12. Results of Standard LightGBM

Split	Accuracy	Precision	Recall	F1-Score	Train Time (s)	Execution Time (s)	Iterations	Avg. Depth	No. of Features
90/10	0.9962	0.9962	0.9962	0.9962	0.4282	0.5766	4,560	3.00	20
80/20	0.9956	0.9956	0.9956	0.9956	0.3786	0.5449	4,560	3.00	20
70/30	0.9954	0.9954	0.9954	0.9954	0.3291	0.4920	4,560	3.00	20

Table 13. Results of Hybrid MVA Based LightGBM

Split	Accuracy	Precision	Recall	F1-Score	Train Time (s)	Execution Time (s)	Iterations	Avg. Depth	No. of Features
90/10	0.9962	0.9962	0.9962	0.9962	0.3543	0.4951	2,964	3.00	13
80/20	0.9956	0.9956	0.9956	0.9956	0.2984	0.4542	2,964	3.00	13

70/30	0.9954	0.9954	0.9954	0.9954	0.2805	0.4381	2,964	3.00	13

Table 14. Error Analysis of Standard and Hybrid MVA Models

Model	Total Test Samples	Correct Predictions	Misclassified	Error Rate (%)	Accuracy
Standard Extra Trees	10,000	9,853	147	1.47	0.9853
Hybrid MVA Extra Trees	10,000	9,928	72	0.72	0.9928
Standard XGBoost	10,000	9,956	44	0.44	0.9956
Hybrid MVA XGBoost	10,000	9,956	44	0.44	0.9956
Standard LightGBM	10,000	9,956	44	0.44	0.9956
Hybrid MVA LightGBM	10,000	9,956	44	0.44	0.9956

Table 15. Cross-Validation Comparison of Standard and Hybrid MVA Models

Model	Features	5-Fold Accuracy (Mean ± SD)	10-Fold Accuracy (Mean ± SD)	5-Fold F1 (Mean ± SD)	10-Fold F1 (Mean ± SD)
Extra Trees	20	0.9849 ± 0.0025	0.9867 ± 0.0017	0.9849 ± 0.0025	0.9867 ± 0.0017
Hybrid MVA Extra Trees	13	0.9923 ± 0.0005	0.9923 ± 0.0013	0.9923 ± 0.0005	0.9923 ± 0.0013
XGBoost	20	0.9954 ± 0.0004	0.9954 ± 0.0009	0.9954 ± 0.0004	0.9954 ± 0.0009
Hybrid MVA XGBoost	13	0.9954 ± 0.0004	0.9954 ± 0.0009	0.9954 ± 0.0004	0.9954 ± 0.0009
LightGBM	20	0.9954 ± 0.0004	0.9954 ± 0.0009	0.9954 ± 0.0004	0.9954 ± 0.0009
Hybrid MVA LightGBM	13	0.9954 ± 0.0004	0.9954 ± 0.0009	0.9954 ± 0.0004	0.9954 ± 0.0009

Table 16. Statistical Significance Analysis

Algorithm	CV	Standard Accuracy	MVA Accuracy	Improvement	Wilcoxon (W)	(p)-value	Significant after Bonferroni?
Extra Trees	5-Fold	0.9849	0.9923	+0.0074	0	0.06250	**No**
Extra Trees	10-Fold	0.9867	0.9923	+0.0056	0	**0.00195**	**Yes**
XGBoost	5-Fold	0.9954	0.9954	0.0000	0	1.00000	**No**
XGBoost	10-Fold	0.9954	0.9954	0.0000	0	1.00000	No
LightGBM	5-Fold	0.9954	0.9954	0.0000	0	1.00000	No
LightGBM	10-Fold	0.9954	0.9954	0.0000	0	1.00000	No

Table 17. SHAP-Based Feature Selection Results

Model	Split	Accuracy	Precision	Recall	F1	Train Time	Execution Time	Iterations	Avg. Depth	Features
SHAP + Extra Trees	90/10	0.8170	0.8180	0.8170	0.8161	0.5261	0.0707	3952	4.0	13
SHAP + XGBoost	90/10	0.9962	0.9962	0.9962	0.9962	0.3097	0.0126	2964	3.0	13
SHAP + LightGBM	90/10	0.9962	0.9962	0.9962	0.9962	0.2563	0.0113	3952	4.0	13
SHAP + Extra Trees	80/20	0.8227	0.8243	0.8227	0.8217	0.4530	0.0728	3952	4.0	13
SHAP + XGBoost	80/20	0.9956	0.9956	0.9956	0.9956	0.2712	0.0141	2964	3.0	13

SHAP + LightGBM	80/20	0.9956	0.9956	0.9956	0.9956	0.2242	0.0164	3952	4.0	13
SHAP + Extra Trees	70/30	0.8415	0.8418	0.8415	0.8410	0.4372	0.0746	3952	4.0	13
SHAP + XGBoost	70/30	1.0000	1.0000	1.0000	1.0000	0.2857	0.0167	2964	3.0	13
SHAP + LightGBM	70/30	0.9954	0.9954	0.9954	0.9954	0.2189	0.0203	3952	4.0	13

Table 18. Overall Result of Proposed Techniques

Models	**Average Accuracy**	**Average Precision**	**Average Recall**	**Average F1**	**Average Iterations**
Extra Trees	0.9859	0.9859	0.9859	0.9859	46,786.67
Hybrid MVA Extra Trees	0.9907	0.9907	0.9907	0.9907	31,039.67
XGBoost	0.9973	0.9973	0.9973	0.9973	4,560
Hybrid MVA XGBoost	0.9973	0.9973	0.9973	0.9973	2,964
LightGBM	0.9957	0.9957	0.9957	0.9957	4,560
Hybrid MVA LightGBM	0.9957	0.9957	0.9957	0.9957	2,964

Performance Analysis of Standard and MVA-Based Extra Trees Models:

The experimental results in table 8 and 9 shows that the Hybrid MVA Extra Trees model has better performance as compared to standard Extra Trees classifier especially in improving computational efficiency. In particular the standard Extra Trees model has accuracies ranging from 98.53% to 98.64% whereas the hybrid MVA model with accuracies ranging from 98.99% to 99.13% which shows that improvement in classification accuracy is very small. A major advantage of the hybrid MVA extra trees is the reduction in computational complexity. The number of input features is reduced from 20 to 13 which means that the number of iteration steps decreases from 47560 to 31655, 46460 to 30862 and 46340 to 30602 across the three experimental splits. Consequently, both training time and execution time are also minimized while maintaining competitive predictive accuracy. Overall, these results proves that the hybrid MVA extra trees feature selection strategy effectively removes less informative attributes which leads to a more efficient model without having any effect on classification performance. This makes it more suitable for real world medical decision support systems where computational efficiency is very important.

Performance Analysis of Standard and MVA XGBoost Models:

The results in table 10 and 11 shows that the hybrid MVA XGBoost model achieved slightly better performance than the standard XGBoost model in terms of computational efficiency across all train test splits. The hybrid MVA approach reduced the number of features from 20 to 13 i.e 35% features reduction and decreased the number of iterations from 4560 to 2964 which means 35% reduction. Furthermore, training and execution time were also reduced across all three train-test splits. The results shows that the hybrid MVA effectively removes redundant features and enhances the computational efficiency of the XGBoost classifier while maintaining its predictive capability.

Performance Analysis of Standard and MVA LightGBM Models:

The results in table 12 and 13 shows that the hybrid MVA LightGBM model consistently outperformed the standard LightGBM classifier in terms of computational efficiency across all three train and test splits. Similar results were observed in terms of accuracy, precision, F1 score and recall. The hybrid MVA approach reduced the number of features from 20 to 13 i.e 35% features reduction and decreased the number of iterations from 4560 to 2964 which means 35% reduction. Furthermore, training time decreased from 0.4282 s to 0.3543 s, 0.3786 s to 0.2984 s and 0.3291 s to 0.2805 s across the three train-test splits. Execution time was also consistently lower in hybrid MVA LightGBM than that of the standard LightGBM model. These results demonstrate that MVA effectively removes redundant features resulting in a faster and more efficient LightGBM model without compromising classification performance.

According to the table 14, the error analysis shows that reducing features in the Hybrid MVA approach gives better performance of the Extra Trees classifier. The number of misclassified samples decreases from 147 in Standard Extra Trees to 72 in case of Hybrid MVA Extra Trees which reduces the error rate from 1.47% to 0.72%. In case of XGBoost and LightGBM, both the standard and Hybrid MVA models have achieved the same performance with 44 misclassified samples and an error rate of 0.44%.

According to table 15, the cross-validation results of all the six models have been carried out. Hybrid MVA Extra Trees improves accuracy by 0.74% at 5-fold from 0.9849 ± 0.0025 to 0.9923 ± 0.0005 and 0.56% at 10-fold while also reducing the feature set from 20 to 13. The Hybrid MVA models of XG Boost and LightGBM achieved the same cross validation accuracy as their standard counterparts with 0.9954 ± 0.0004 in 5 fold and 0.9954 ± 0.0009 in 10 fold validation. This shows that by the elimination of seven features, the predictive performance did not degrade in either boosting model. It also shows that the model performance has been consistent due low standard deviations across the folds. The

findings of table 15 also supports the robustness and effectiveness of the proposed hybrid mva feature selection method.

Statistical Significance Analysis:

A paired Wilcoxon signed rank test was performed to identify the performance differences between standard models and their hybrid MVA models. The test was evaluated for both models on the same cross validation folds. For each fold, the difference between the MVA and standard model was calculated as follows:

$$di=AiMVA - Ai\ Standard \qquad 10$$

where the equation 11 is the performance of hybrid MVA and standard models on fold (i). The absolute non zero differences were ranked and the positive and negative rank sums were calculated as:

$$W^{+}=\sum_{di>0} Ri \text{ and } W^{-}=\sum_{di<0} Ri \qquad 11$$

where Ri denotes the rank assigned to the absolute value of the (i)-th paired difference The Wilcoxon test statistic was then obtained as:

$$W=\min (W^{+}, W^{-}) \qquad 12$$

This statistic is subsequently used to determine the corresponding (p) value. The hypotheses were defined as:

The null hypothesis H_0 as there is no systematic performance difference between the standard and MVA models whereas the alternative hypothesis H_1 states that a systematic difference exists. Statistical significance was evaluated at alpha=0.05 with $p<0.05$ indicating a statistically significant difference. Since six standards versus MVA comparisons were considered, a Bonferroni correction was applied:

$$\alpha_{adjusted}=0.05/6=0.00833 \qquad 13$$

Therefore, $p<0.00833$ was considered statistically significant after multiple-comparison correction. The resulting Wilcoxon W statistics and p values are reported in Table 10 below.

As per the equation 14, the bonferroni threshold is 0.00833 so if the value of P is less than 0.00833 it means that the mode is statistically significant and if the value of P is greater than 0.00833 then it means that the model is not significant. According to the table 16 only the 10-fold ET comparison is statistically significant as $p=0.00195<0.00833$. The 5 fold Extra Trees comparison and both XGBoost and LightGBM comparisons did not show statistically significant differences after Bonferroni correction.

SHAP integration with ensemble models:

There are several feature selection methods that were used such as ReliefF, Boruta, mRMR and SHAP. However, SHAP was selected as it provides a model-based explanation of feature importance instead of just simply measuring the features through their statistical relationships. According to table 17, the SHAP approach has selected 13 features for all three train test splits and the results shows that SHAP based Extra Trees achieved accuracy between 81.70% and 84.15% across the different splits which is lower than standard extra trees. In contrast SHAP based XGBoost and LightGBM produced substantially higher performance. XGBoost also required fewer iterations with an average depth of 3 compared while LightGBM showed the lowest training time in all three splits.

The proposed Hybrid MVA framework as compared with the SHAP based approach has achieved a better balance between feature reduction, predictive accuracy and computational efficiency. It also provides an outstanding predictive performance as compared to SHAP in case of extra trees. Moreover, it provides an important advantage over SHAP because feature selection is based on the combination of MVA cardinality and variance rather than relying on a model specific explanation method. The comparison demonstrates that the improvement is not simply due to using fewer features but is associated with the proposed MVA based selection strategy.

Comparative Discussion:

The experimental results shown above of standard classifiers and hybrid MVA based classifiers demonstrates that the hybrid MVA models effectively reduced computational cycles and manage predictive performance comparable to their standard classifiers. The hybrid MVA approach reduces the input features from 20 to 13 before model training. In the standard extra trees, the iteration steps were 46786.67 however in hybrid MVA model it has been reduced to 31039.67 i.e almost 33.66% reduction in computational complexity. Similarly, training and execution time were also reduced in hybrid MVA Extra trees. The Hybrid MVA XG Boost model and hybrid MVA Light GBM reduced the average iteration steps from 4560 to 2964 which is almost a 35% reduction and also reducing training and execution times.

According to table 18, the average performance Hybrid MVA classifiers indicates that they reduce the number of selected features and iteration steps while maintaining classification performance comparable to the standard models. Among the evaluated classifiers, the hybrid MVA XGBoost gives the highest average predictive performance while reducing iteration steps and number of features. Hybrid MVA LightGBM also maintain predictive performance and reduces iteration steps and features. The Hybrid MVA Extra Trees model also maintain prediction accuracy while substantially reducing computational overhead compared with the standard Extra Trees classifier, however the number of iteration steps in Hybrid MVA extra trees is still large as compared to Hybrid MVA XG Boost and hybrid MVA light GBM. These results validate the effectiveness of the proposed Hybrid MVA feature selection framework for heart disease prediction.

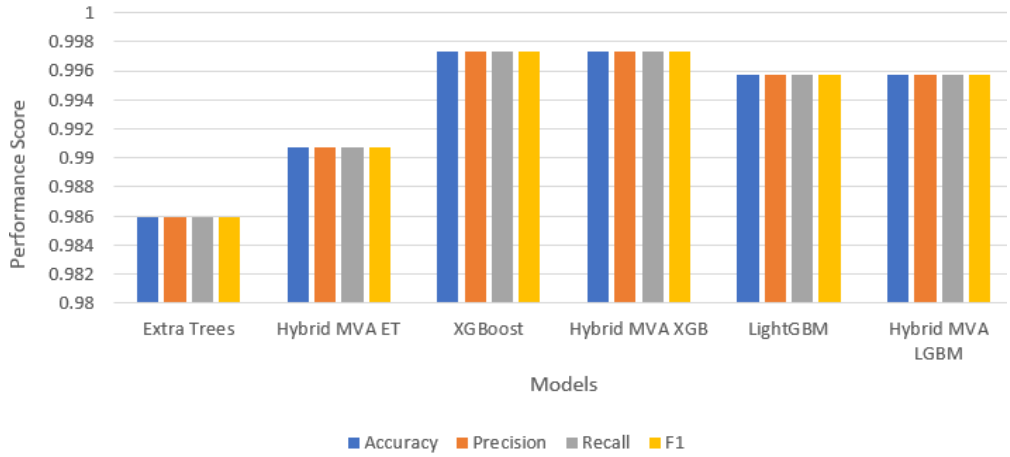

Figure 3. Average Classification Performance of Standard and Hybrid MVA Models

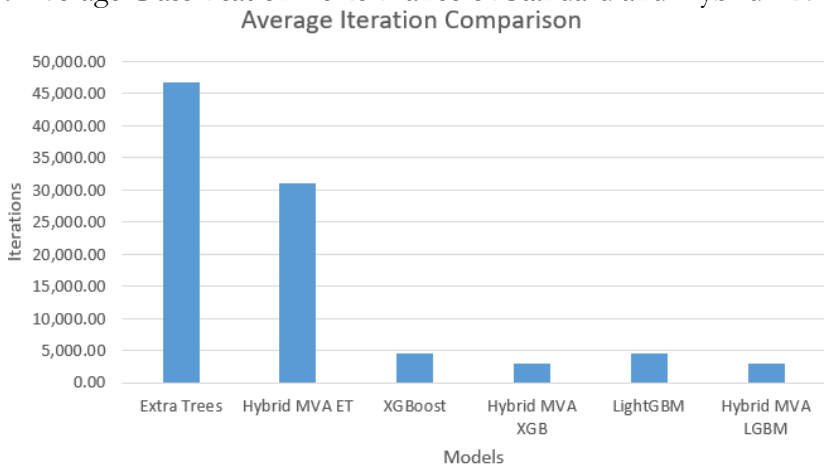

Figure 4. Average Iteration Comparison

Key Findings:

Practical Deployment:

The experimental results shows that the proposed Hybrid MVA has potential to use in clinical decision support applications. As we saw that the proposed approach not only reduces the original feature space to 13 selected attributes but also maintains strong predictive performance. This reduction could be very useful in many clinical systems because the model needs to process only the relevant attributes for every patient for prediction. This will also reduce the computational workload during model inference so the trained models can generate predictions within a short period of time. Therefore, in clinical decision support systems where predictions may need to be generated immediately after patient information becomes available. The relatively low execution time observed in the experiments is more relevant to real time usage than the training time. The results also show that Hybrid MVA across different train and test splits is computationally manageable and the number of selected features remains fixed at 13 whereas the models were executed with controlled model depth.

Scalability:

As in hybrid MVA, each prediction is performed using a reduced set of attributes, therefore the computational workload with each patient records will be very limited. The proposed approach could be incorporated into a clinical decision support system where patient records are achieved from the clinical information system. In the current experiments the required 13 attributes are extracted from research dataset and the trained model then produces the prediction, so the current experiments do not represent an actual hospital deployment. Future work may implement the framework using larger clinical datasets and assess its performance under real clinical workloads such as simultaneous requests, hardware specific latency, data availability and integration with electronic health systems.

MVA based models improved prediction performance:

The experimental results demonstrate that the proposed hybrid MVA ensemble models achieved better heart disease prediction performance compared to the conventional Extra Trees, XGBoost and LightGBM models.

Reduction in training time:

The integration of the hybrid MVA technique reduced the overall training time of the ensemble learning models.

Reduction in model complexity:

The proposed hybrid MVA based models reduced model complexity by optimizing feature selection and minimizing redundant computational operations.

Reduced iteration and computational steps:

The proposed approach reduced the number of computational and iteration steps that are required for model training and prediction.

Improved accuracy and classification metrics:

The hybrid MVA based ensemble models achieved almost similar classification performance in terms of accuracy, precision, recall and F1 score.

Hybrid MVA XG Boost performed best:

Among all evaluated models Hybrid MVA XGBoost achieved the best overall predictive performance on the disease prediction dataset in terms of predictive performance, iteration steps, training and execution time.

Limitations of the Study:

The main limitation of this study is the use of Kaggle dataset rather than real world clinical data which could be collected directly from patients. Although the dataset provides important details for evaluating the proposed Hybrid MVA framework but it may not fully represent the complexity and variability present in actual clinical records. Real clinical

datasets can contain missing values, differences in patient characteristics and differences in collection of data practices. All these may not be completely reflected in a synthetic Kaggle dataset. Another important limitation is generalizability because in this study, the models are trained and evaluated on only one dataset, therefore it may perform differently when applied to patients from different hospitals, locations or healthcare settings. Future work should therefore validate the proposed Hybrid MVA framework on real world clinical datasets obtained from multiple healthcare institutions.

Conclusion:

The proposed Integrated Predictive Model can be helpful for health management authorities as integrating hybrid MVA based feature selection with advanced algorithms like XGBoost and Extra Trees optimizes heart disease prediction efficiency. The experimental finding reveals that while standard models like Extra Trees, XGBoost and LightGBM provide high accuracy but they are often hindered by excessive iteration steps and execution time. In contrast the hybrid MVA integrated models such as hybrid MVA ET, hybrid MVA XGBoost and hybrid MVA LightGBM significantly reduce computational cycles, execution time, training time and maintain predictive performance as compared to standard models.The hybrid MVA XGBoost offered the best tradeoff between predictive performance and computational efficiency. The hybrid MVA XGBoost lowers the iteration steps from 4560 to 2964 corresponding to a 35% reduction and also reduces the training and execution time which is a big difference. These integrated models are highly viable for large scale healthcare setups requiring prompt medical decision making. The integration of RST through proposed hybrid MVA provides a viable solution for real time medical diagnostic tools where early and timely detection is paramount for improving early diagnosis and optimizing healthcare resources.

Future Work:

There are following directions for future research that are identified to increase prediction accuracy or computational cost of classifiers in terms of health diseases:

Clinical Validation on Real Healthcare Data: The proposed hybrid MVA method can also be implemented using real world clinical datasets collected from hospitals and healthcare organizations to validate its effectiveness under practical medical conditions.

Expansion to New Viral Variants and Diseases: The current MVA based framework has the potential to be adapted for other threatening and deadly diseases for e.g the Congo virus, Zika and emerging variants of COVID19 like Omicron or BF7.

Deep Learning Comparisons: To inherent account for temporal dynamics in disease outbreaks future studies could compare the performance of MVA models against Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) architectures.

Clinical Deployment: A primary goal for future work is the transition from research simulations to real time mobile diagnostic applications in order to reduce computational cycles of MVA and to allow low-cost screening in resource limited regions.

References:

[1] K. Gulzar, B. A. Hassan, M. Abbas, Z. Ahmad, and M. A. and H. Ullah, "Maximum Value Attribute based Decision Tree and Random Forest for COVID-19 Prediction," *Int. J. Innov. Sci. Technol.*, vol. 7, no. 4, pp. 2940–2954, 2025, Accessed: Aug. 17, 2026. [Online]. Available: https://ideas.repec.org/a/abq/ijist1/v7y2025i4p2940-2954.html

[2] Das, P.; Sarker, P.; Tiang, J.-J.; Nahid, A.-A., "Dengue Fever Classification Integrating Bird Swarm Algorithm with Gradient Boosting Classifier Along with Feature Selection and SHAP–DiCE Based Interpretability.," *Appl. Sci*, vol. 15, no. 11413, 2025, doi: https://doi.org/10.3390/app152111413.

[3] I. \Izaz Ahmmed Tuhin, A.K.M.Fazlul Kobir Siam, Md Mahfuzur Rahman Shanto, Md Rajib Mia, "'An interpretable machine learning model for dengue detection with

clinical hematological data,"' *Healthc. Anal.*, vol. 8, 2025, doi: https://doi.org/10.1016/j.health.

[4] T. Ashika and G. H. Grace, "Enhancing Classification Performance through Rough Set Theory Feature Selection: A Comparative Study across Multiple Datasets," *Eur. J. Pure Appl. Math.*, vol. 18, no. 2, Apr. 2025, doi: 10.29020/NYBG.EJPAM.V18I2.5934.

[5] T. Ashika and G. Hannah Grace, "Enhancing heart disease prediction with stacked ensemble and MCDM-based ranking: an optimized RST-ML approach," *Front. Digit. Heal.*, vol. 7, p. 1609308, Jun. 2025, doi: 10.3389/FDGTH.2025.1609308/TEXT.

[6] D. Dewi, B.E.; Kartika, A.A.A.; Faridah, A.T.; Ewaldo, M.F.; Hafizh, A.M.; Chrysilla, V.; Frederich, J.; Surya, A.; Aryani, "Development of a Machine Learning Model for Predicting Dengue Cases and Severity in Indonesia,"," *Appl. Sci*, vol. 16, no. 1436, 2026, doi: https://doi.org/10.3390/app16031436.

[7] A. M. Khan S, Ullah R, Khan A, Wahab N, Bilal M, ""Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM)," *Biomed Opt Express*, vol. 7, p. 2249, 2016, doi: doi: 10.1364/BOE.7.002249.

[8] O. E. Kumar Y, Liang C, Bo Z, Rajapakse JC, ""Serum Proteome and Cytokine Analysis in a Longitudinal Cohort of Adults with Primary Dengue Infection Reveals Predictive Markers of DHF," *PLoSNegl Trop Dis*, vol. 6, 2021, doi: doi: 10.1371/journal.pntd.0001887.

[9] N. Zhao et al., "'Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia,"' *," PLoSNegl Trop Dis*, vol. 14, pp. 1–16, 2020, doi: , doi: 10.1371/journal.pntd.0008056.

[10] S. Subudhi *et al.*, "Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19," *npj Digit. Med. 2021 41*, vol. 4, no. 1, pp. 87-, May 2021, doi: 10.1038/s41746-021-00456-x.

[11] D. O. Uchenna J. Nzenwata1* , Emokiniovo Edwin1, Emmanuel A. Chukwu2 and C. E. Johnson O. Hinmikaiye1, "Extra Trees Model for Heart Disease Prediction," *J. Data Anal. Inf. Process.*, vol. 13, p. . 125–139, 2025, doi: https://doi.org/10.4236/jdaip.2025.132008.

[12] B. Liu, M. F. Hossain, and S. Hossain, "A comparative evaluation of multiple machine learning approaches for forecasting dengue outbreaks in Bangladesh," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 35931-, Oct. 2025, doi: 10.1038/s41598-025-19752-7.

[13] M. A. & R. D. . Bam Bahadur Sinha, "LightGBM empowered by whale optimization for thyroid disease detection,"," *Int. J. Inf. Technol. (Singapore),* vol. 15, pp. 2053–2062, 2023, doi: doi.org/10.1007/s41870-023-01261-3.

[14] J. Uddin, R. Ghazali, M. M. Deris, U. Iqbal, and I. A. Shoukat, "A novel rough value set categorical clustering technique for supplier base management," *Comput. 2021 1039*, vol. 103, no. 9, pp. 2061–2091, Apr. 2021, doi: 10.1007/S00607-021-00950-W.

[15] A. Khakharia *et al.*, "Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning," *Ann. Data Sci. 2020 81*, vol. 8, no. 1, pp. 1–19, Oct. 2020, doi: 10.1007/S40745-020-00314-9.

[16] W. M. Alhadi Bustamam 1,*, Hengki Muradi 2, W. Mangunwardoyo, S. orgGoogle. Scholar, and 3 andBeti E. Dewi 4, "Comparison of dengue predictive models developed using artificial neural network and discriminant analysis with small dataset," *Appl. Sci.*, vol. 11, no. 3, 2021, doi: https://doi.org/10.3390/app11030943.

[17] H. Fayyoumi, E., Idwan, S., & AboShindi, "Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan,"," *Jordan*, vol. 11, no. 5, 2020, doi: DOI: https://doi.org/10.14569/IJACSA.2020.0110518.

[18] M. C. Y. H. K. H. L. W. L. W. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities,"," *IEEE Access*, vol. 5, p. . 8869-8879, 2017, doi: doi: 10.1109/ACCESS.2017.2694446.

[19] E. Arianyan, N. Gholipour, D. Maleki, N. Ghorbani, A. Sepahvand, and P. Goudarzi, "A Systematic Review and Classification of HPC-Related Emerging Computing Technologies," *Electron. 2025, Vol. 14, Page 2476*, vol. 14, no. 12, p. 2476, Jun. 2025, doi: 10.3390/ELECTRONICS14122476.

[20] A. A. Alrajhi *et al.*, "Data-Driven Prediction for COVID-19 Severity in Hospitalized Patients," *Int. J. Environ. Res. Public Health*, vol. 19, no. 5, Mar. 2022, doi: 10.3390/IJERPH19052958.

[21] L. J. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, and I. A. Mohammed, "Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset," *SN Comput. Sci. 2020 21*, vol. 2, no. 1, pp. 11-, Nov. 2020, doi: 10.1007/S42979-020-00394-7.

[22] Y. Zoabi and N. Shomron, "COVID-19 diagnosis prediction by symptoms of tested individuals: a machine learning approach," *medRxiv*, p. 2020.05.07.20093948, May 2020, doi: 10.1101/2020.05.07.20093948.

[23] D. Arumuganainar, "Extra Tree Classifier Predicts an Interactome Hub Gene as HSPB1 in Oral Cancer: A Bioinformatics Analysis," *Cureus*, May 2024, doi: 10.7759/CUREUS.59863.

[24] Y. Gao *et al.*, "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19," *Nat. Commun. 2020 111*, vol. 11, no. 1, pp. 5033-, Oct. 2020, doi: 10.1038/s41467-020-18684-2.

[25] D. K. Sharma, M. Subramanian, P. Malyadri, B. S. Reddy, M. Sharma, and M. Tahreem, "Classification of COVID-19 by using supervised optimized machine learning technique," *Mater. Today Proc.*, vol. 56, pp. 2058–2062, Jan. 2022, doi: 10.1016/J.MATPR.2021.11.388.

[26] M. E. Haque, S. M. Jahidul Islam, J. Maliha, M. S. Hossan Sumon, R. Sharmin, and S. Rokoni, "Improving Chronic Kidney Disease Detection Efficiency: Fine Tuned CatBoost and Nature-Inspired Algorithms with Explainable AI," *2025 IEEE 14th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2025*, pp. 811–818, Apr. 2025, doi: 10.1109/CSNT64827.2025.10968421.

[27] M. N. Islam, M. M. H. Rimon, S. S.-E.-A. Shamim, Z. M. Fahad, M. J. I. Mony, and M. J. U. Chowdhury, "An Improved Ensemble-Based Machine Learning Model with Feature Optimization for Early Diabetes Prediction," Nov. 2025, Accessed: Aug. 18, 2026. [Online]. Available: https://arxiv.org/pdf/2512.02023

[28] Ahmet ÇELİK, "Predicting Diagnosis Of Covid-19 Disease With Adaboost And Naive Bayes Machine Learning Algorithms," *J. Eng. Sci. Des.*, vol. 10, no. 4, 2022, [Online]. Available: https://dergipark.org.tr/en/download/article-file/1901380

[29] "Appositeness of Hoeffding tree models for breast cancer classification | Journal of Current Science and Technology." Accessed: Aug. 18, 2026. [Online]. Available: https://ph04.tci-thaijo.org/index.php/JCST/article/view/253

[30] M. S. Satu *et al.*, "Short-Term Prediction of COVID-19 Cases Using Machine Learning Models," *Appl. Sci. 2021, Vol. 11, Page 4266*, vol. 11, no. 9, p. 4266, May 2021, doi: 10.3390/APP11094266.

[31] A. Z. Baratpur, H. Vahdat-Nejad, E. Arslan, J. Hassannataj Joloudari, and S. Gaftandzhieva, "Coronary artery disease prediction using Bayesian-optimized support vector machine with feature selection," *Front. Netw. Physiol.*, vol. 5, p. 1658470, Dec. 2025, doi: 10.3389/FNETP.2025.1658470/TEXT.

[32] D. N. M. S. R. S. K.Tharageswari, "A Robust Disease Prediction System Using

Hybrid Deep Neural Networks," *Libr. Prog. Int.*, vol. 44, no. 3, pp. 28383–28398, Nov. 2024, doi: 10.48165/BAPAS.2024.44.2.1.

[33] K. Alsharabi, Y. Bin Salamah, A. M. Abdurraqeeb, M. Aljalal, and F. A. Alturki, "EEG Signal Processing for Alzheimer's Disorders Using Discrete Wavelet Transform and Machine Learning Approaches," *IEEE Access*, vol. 10, pp. 89781–89797, 2022, doi: 10.1109/ACCESS.2022.3198988.

[34] A. Alariyibi, M. El-Jarai, and A. Maatuk, "Evaluating The Accuracy of Classification Algorithms for Detecting Heart Disease Risk," *Mach. Learn. Appl. An Int. J.*, vol. 10, no. 4, pp. 01–12, Dec. 2023, doi: 10.5121/mlaij.2023.10401.

[35] M. S. Alkhasawneh, "Hybrid Cascade Forward Neural Network with Elman Neural Network for Disease Prediction," *Arab. J. Sci. Eng. 2019 4411*, vol. 44, no. 11, pp. 9209–9220, Apr. 2019, doi: 10.1007/S13369-019-03829-3.

[36] A. Y. Yıldız and A. Kalayci, "Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data," *2025 IEEE Conf. AI Data Anal. ICAD 2025*, Aug. 2025, doi: 10.1109/ICAD65464.2025.11114069.

[37] Y. On *et al.*, "Classification of Mitral Regurgitation from Cardiac Cine MRI Using Clinically-Interpretable Morphological Features," *Lect. Notes Comput. Sci.*, vol. 15448 LNCS, pp. 245–256, 2025, doi: 10.1007/978-3-031-87756-8_25/SAVE-RESEARCH.

[38] S. Noor, S. A. AlQahtani, S. Khan, S. Noor, S. A. AlQahtani, and S. Khan, "Chronic liver disease detection using ranking and projection-based feature optimization with deep learning," *AIMS Bioeng. 2025 150*, vol. 12, no. 1, pp. 50–68, 2025, doi: 10.3934/BIOENG.2025003.

[39] Y. Hu and A. Chaddad, "SHAP-Integrated Convolutional Diagnostic Networks for Feature-Selective Medical Analysis," Mar. 2025, Accessed: Aug. 17, 2026. [Online]. Available: https://arxiv.org/pdf/2503.08712

[40] B.-C. Phan, T. Ma, H.-H. Nguyen, and T.-N. Do, "BoMGene: Integrating Boruta-mRMR feature selection for enhanced Gene expression classification," Oct. 2025, Accessed: Aug. 17, 2026. [Online]. Available: https://arxiv.org/pdf/2510.00907