Grenze International Journal of Engineering and Technology, June Issue

Optimized Machine Learning Frame Work for Heart Disease Prediction

Dr.U Moulali[1], M. Rithvik Yadav[2], Dasari Srikanth[3] and K.Sandeep Kumar[4]
[1]Assosiate Professor,
[1-4]Department of Computer Science and Engineering,
Methodist College of Engineering and Technology, Hyderabad, India
Email: moulaliu@methodist.edu.in, rithvikyadav040@gmail.com, dasarisrikanth1111@gmail.com,
Kakunurisandeep12@gmail.com

***Abstract*—Globally, cardiovascular illnesses continue to rank among the top causes of death. for an intervention to be effective, risks must be accurately and quickly identified. This research presents an Optimized Machine Learning Framework aimed at predicting heart disease, which integrates traditional Machine Learning models with the WOA to boost precision in forecasting. For our analysis we made use of the C level and Heart Disease Data set and the Healthcare Stroke Dataset, to enhance the model's reliability and overall applicability. To address the class imbalance in the Healthcare Stroke Dataset, SMOTE was utilized, followed by choosing relevant features using SHAP. We pinpointed the five most important features affecting the results and incorporated them into training the model with Logistic Regression, Random Forest and XGBoost. Of these methods, XG Boost attained thehighestaccuracyat86%.Further adjustments using WOA increased the accuracy to 93%, demonstrating the algorithm's effectiveness in fine-tuning model parameters. Furthermore, we used SHAP to improve the model's clarity and assess the significance of features. In the analysis of the Cleveland Heart Disease Dataset, we implemented a similar training approach with the same algorithms, resulting in Logistic Regression achieving an accuracy of 86%. Subsequent optimization with WOAled to an increase to 87%. By utilizing permutation-based feature importance, we were able to recognize key predictive elements. The Healthcare Stroke Dataset mainly emphasizes detecting stroke risk factors, whereas the Cleveland Heart Disease Dataset is centered on forecasting heart illnesses. By evaluating the model's performance on both datasets, we can assess the framework's effectiveness across different yet related cardiovascular issues, thereby enhancing its usefulness and applicability in real-world medical contexts.**

***Index Terms*— Cardiovascular illnesses, WOA (WhaleAlgorithm), SMOTE (Synthetic Optimization Minority Over-sampling Technique),SHAP(Shapley Additive Explanations).**

I. INTRODUCTION

An estimated 17.9 million deaths worldwide are attributed to cardiovascular diseases, making them the top cause of death worldwide. [1][3][4].It is crucial to identify risks correctly and promptly for effective intervention. Traditional diagnostic techniques, such as electro cardiograms, echo cardiograms, and stress tests, require clinical expertise and often fail to leverage the vast amount of available patient data for early risk assessment[3]. Machine learning (ML) has emerged as a powerful tool to enhance heart disease prediction by identifying

Grenze ID: 01.GIJET.11.2.606_23

complex patterns in patient data that may not be evident through conventional methods. However, challenges such as imbalanced datasets, irrelevant feature selection, and hyper parameter tuning influence ML models' generalizability and performance in applications real-world [2].Recent studies clinical has concentrated on using a variety of classification algorithms, such as logistic regression, support vector machines (SVM), random forests, and deep learning-based techniques, to optimize machine learning models for the prediction of heart disease[1][3]. The selection of appropriate features and tuning of hyperparameters significantly influence the predictive accuracy of these models. Hyperparameter tuning, in particular, is a criticalaspectofMLmodeldevelopment,asitdirectlyimpactsthemodel’sabilitytogeneralizeacrossdifferent datasets and patient demographics[2]. Automated hyperparameter optimization techniques have been introduced to fine-tune model parameters efficiently, leading to improved classification performance and reduced computational cost[7] .In this research, an optimized ML framework is proposed for for predicting cardiacdiseasebyintegratingtraditionalclassificationalgorithmswithmetaheuristicoptimization technique. Specifically, the incorporates Algorithm the Whale framework Optimization to fine-tune hyperparameters, thereby improving model efficiency and robustness. Additionally, SHAP is employed to enhance the interpretability of predictions, making the results more accessible to healthcare professionals. To assess the effectiveness of the proposed method, two benchmark datasets the Cleveland heart disease dataset and the Healthcare stroke data set are used. The Healthcare Stroke Dataset primarily focuses on detecting stroke risk factors, with SHAP used for feature selection. The Cleveland Heart Disease Dataset is centered on predicting heart disease risk, where permutation-based feature importance is employed to identify key predictors.

II. LITERATURE REVIEW

In this research, a proposed hyper parameter tuning system is compared to a conventional heart disease prediction system. To determine the ideal hyperparameters, the suggested method uses a Grid search strategy in conjunction with five machine learning algorithms. The suggested hyperparameter tuning model produced accuracies between 85.25% and 91.80%, whereas the conventional method produced results between 81.97% and90.16%[1].This research investigates the prediction of strokes using neural networks and machine learning by analyzing electronic health records (EHR) and identifying key risk factors such as age, heart disease, hypertension, and average glucose levels, while employing Principal Component Analysis (PCA) to reduce data dimensionality[2].This research involves applying various machine learning classification techniques in orderto forecast cardic illness.It employs algorithms such as decision tree,k-nearest neighbors, random forest ,support vector machine, multilayer perceptron, and XGBoost. Additionally, k-modes clustering with Huang initialization is utilized to enhance classification accuracy[3]. Cardiovascular disease prediction helps practitioners make accurate health decisions, enabling early detection for lifestyle changes and timely medical intervention. Machine learning (ML) is utilized to analyze heart disease symptoms, with the chi-square test picking important characteristics from the Cleveland dataset on cardiac disease. Several machine learning techniques are used, with accuracies ranging from 73.77% to 88.5%. These algorithms include svm, gaussian naive bayes, logistic regression, light GBM, XGBoost, and random forest. During validation on 303 instances with 13 chosen features, the Random Forest algorithm shows the greatest accuracy of 88.5%, aided by data visualization to examine feature correlations[4].

III. PROPOSEDMETHODOLOGY

A. Architecture

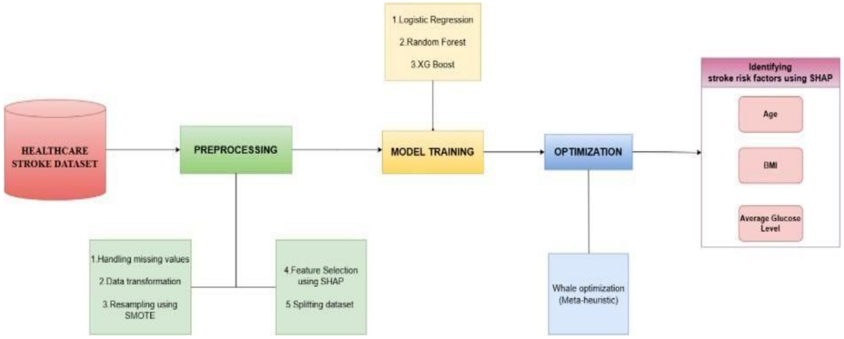


Fig1:Optimized stroke risk assessment model

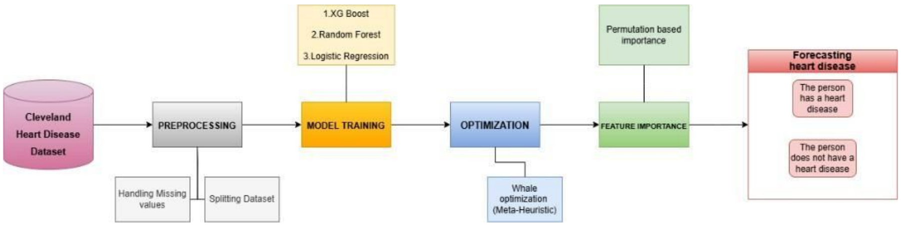

1.XG Boost
2.Random Forest
3.Logistic Regression
Permutation based importance
Cleveland Heart Disease Dataset
PREPROCESSING
MODEL TRAINING
OPTIMIZATION
FEATURE IMPORTANCE
Forecasting heart disease
The person has a heart disease
The person does not have a heart disease
Handling Missing values
Splitting Dataset
Whale optimization (Meta-Heuristic)

Fig2:Heart disease forecasting and optimization framework

The proposed architectures for stroke (fig1)and heart disease prediction (fig2)utilize machine learning models enhanced with WOA to achieve improved accuracy and reliability. Both models follow a structured pipeline involving data preprocessing, model training, optimization, and feature importance analysis. Both architectures demonstrate the effectiveness of WOA in optimizing hyperparameters, resulting in enhanced model interpret ability performance. The offered by SHAP and permutation-based importance ensures that the models are not only accurate but also explainable, making them suitable for real world clinical applications

IV. DATASETDESCRIPTION

TABLE1:HEALTHCARESTROKEDATASET-FEATUREDESCRIPTION

Feature	Type	Description
id	Integer	Uniqueidentifierforeachpatient
gender	Categorical(Male,Female,Other)	Patient'sgender
age	Float	Ageofthepatient
hypertension	Binary(0=No,1=Yes)	Indicatesifthepatienthashighbloodpressure
heart_disease	Binary(0=No,1=Yes)	Indicatesifthepatienthasahistoryofheartdisease
ever_married	Categorical(Yes,No)	Maritalstatusofthepatient
work_type	Categorical(Private,Self-employed,Govt_job,Children,Never_worked)	Typeofemployment
Residence_type	Categorical(Urban,Rural)	Whetherthepatientlivesinanurbanorruralarea
avg_glucose_level	Float	Averageglucoselevelofthepatient
bmi	Float (Somemissingvalues)	BodyMass Index(BMI)
smoking_status	Categorical(formerly smoked,neversmoked,smokes,Unknown)	Smokinghistory
stroke	Binary(0=No,1=Yes)	Targetvariableindicating whetherthepatienthadastroke

TABLE II:C LEVEL AND HEART DISEASE DATASET-FEATURE DESCRIPTION

Feature	Type	Description
age	Integer	Ageofthepatientin years
sex	Binary(1=Male,0=Female)	Genderofthepatient
cp	Categorical(0,1,2,3)	ChestPainType:0:Typicalangina,1:Asymptomatic,2:Non-anginal,3:Nontypicalangina
RestBP	Integer	Restingbloodpressure inmmHg
Chol	Integer	Serumcholesterolinmg/dl
Fbs	Binary(1=True,0=False)	Fastingbloodsugar>120mg/dl
RestECG	Categorical(0,1,2)	RestingECGresults:0:Normal,1:ST-Twaveabnormality,2:LV hypertrophy
MaxHR	Integer	Maximumheartrateachievedduringstresstest
Exang	Binary(1=Yes,0=No)	Exercise-inducedangina
oldpeak	Float	STdepression inducedbyexerciserelativetorest
slope	Categorical(1,2,3)	SlopeofpeakexerciseSTsegment:1:Upsloping,2:Flat,3:Downsloping
ca	Integer(0-3)	Numberofmajorvesselscoloredby fluoroscopy
target	Binary(0=Nodisease,1=Presenceofdisease)	Heartdiseasestatus
thal	Categorical(1,2,3)	Thaliumstresstestresult:1:Fixed,2:Normal,3:Reversibledefect

V. DATA PREPROCESSING

In the Healthcare Stroke Dataset, 201 missing values in the BMI column were addressed using group-wise mean imputation—calculating separate means for stroke (bmi_stroke_mean) and non-stroke (bmi_no_stroke_mean) patients. This method maintained the integrity of the data distribution and prevented bias,recognizing BMI as assign if icantstroker isk factor. In the C level and Heart Disease Dataset, four missing valuesinthe'ca'column were handled with standard mean imputation.Data transformation steps included label encoding of categorical variables suchas'ever_married','Residence_type',and'gender',andone-hotencoding for 'work_type' and 'smoking_status'. Numerical features like 'age', 'bmi', and 'avg_glucose_level' were standardized using Standard Scaler to ensure model compatibility. For the Cleveland dataset, label encoding was applied to categorical attributes like"sex","cp","fbs","restecg","exang","slope","thal",and"ca".Class imbalance in the stroke dataset was mitigated using SMOTE, while the Cleveland dataset remained balanced and required no resampling. Feature selection in the Stroke dataset was conducted using SHAP (SHapley Additive exPlanations),leveragingtheKernelExplainertoidentifythetopfivemostimpactfulfeatures—age, BMI, average glucose level, smoking_status_never smoked, and Residence_type. These were selected for model training to reduce dimensionality while preserving predictive power. SHAP was also used for interpretability, assigning Shapley values to features to explain their contributions to individual predictions. In contrast, the Cleveland dataset employed permutation-based feature importance, measuring the performance drop upon random shuffling of each feature. Partial Dependence Plots (PDPs) further illustrated how specific features influenced predictions. Machine learning models such as XGBoost, SVM, KNN, Random Forest, and Logistic Regression were trained, with optimization using the Whale Optimization Algorithm. Evaluation metrics included accuracy, precision, recall, F1-score, confusion matrix, and model explainability via SHAP.

V. RESULTS AND DISCUSSION

TABLEIII. PERFORMANCECOMPARISONOFDIFFERENTMLMODELSONHEALTHCARESTROKEDATASET

ModelName	Trainaccuracy(%)	Testaccuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
LogisticRegression	78	76	99	76	86
RandomForest	85	74	99	74	84
XGBoost	94	86	98	88	93
KNN	100	83	98	84	91
SVM	80	74	88	71	78

With an 86% test accuracy, XGBoost outperformed the other models during evaluation.Due to its superior performance,we selected XGBoostfor furtheroptimizationusing theWhale OptimizationAlgorithm(WOA) to enhance its predictive capability

TABLEIV: OPTIMIZED XGBOOST HYPER PARAMETERS AND PERFORMANCE USING WHALE OPTIMIZATION ALGORITHM

ModelName	BestParameters(OptimizedwithWOA)	FinalAccuracy(%)	Precision (%)	Recall(%)	F1-Score(%)
XGBoost	n_estimators=169,max_depth=6,learning_rate =0.194,subsample=0.58	93	97	96	96

By applying the WOA, the XGBoost classifier was fine-tuned to achieve better generalization and higher accuracy of 93%, making it a strong candidate for identifying influential factors for cardiovascular disease prediction in this study.

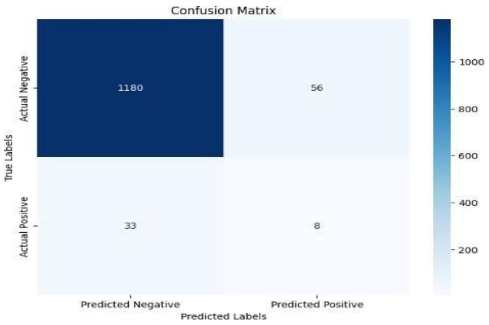


Fig3:Confusion matrixforoptimizedXGBoostwith WOA

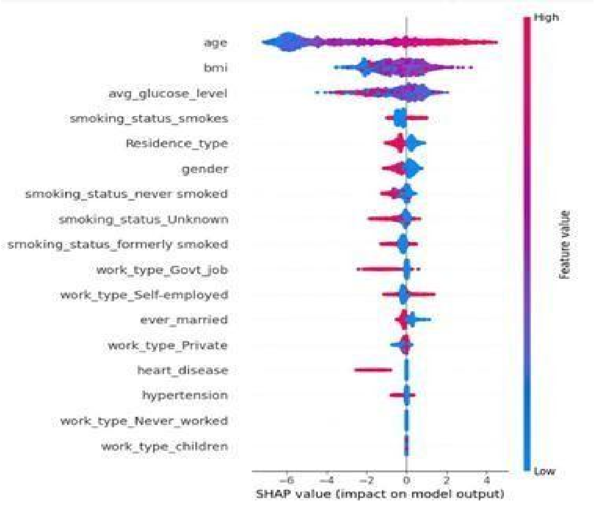


Fig4:Shapsummaryplotforfeatureimportance

The SHAP feature importance plot (fig 4)provides a clear understanding of the factors that most influence cardiovascular disease prediction in the model. Age is the most dominant predictor, significantly impacting modeloutcomes

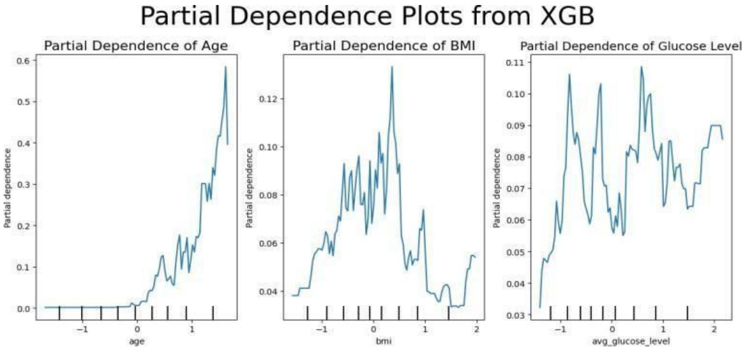


FIG5: Partial dependence plots of key features in XGBoost model

From fig 5 age is the most influential feature among the three, with a clear increasing trend.BMI and avg_glucose_level show smaller and more complex contributions.These plots help interpret how the model uses each feature, improving model transparency and trustworthiness.

TABLEV: PERFORMANCE COMPARISON OF DIFFERENT MLMODELS ON C LEAVEL AND HEART DISEASE DATASET

ModelName	Trainaccuracy(%)	Testaccuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
LogisticRegression	85	86	87	85	86
Random Forest	100	84	84	85	85
XGBoost	100	80	83	79	81
KNN	72	51	53	52	53
SVM	65	67	86	81	78

With a test accuracy of 86%, Logistic regression outperformed the other models under evalua tion. Due to its superior performance, we selected Logistic Regression for further optimization using the Whale Optimization Algorithm (WOA) to enhance its predictive capability.

TABLEVI:OPTIMIZED LOGISTIC REGRESSION HYPER PARAMETERS AND PERFORMANCE USING WHALE OPTIMIZATION ALGORITHM

ModelName	BestParameters(OptimizedwithWOA)	FinalAccuracy (%)	Precision (%)	Recall(%)	F1-Score(%)
LogisticRegression	C=0.4514,Solver=liblinear,Penalty=l2	87	88	88	88

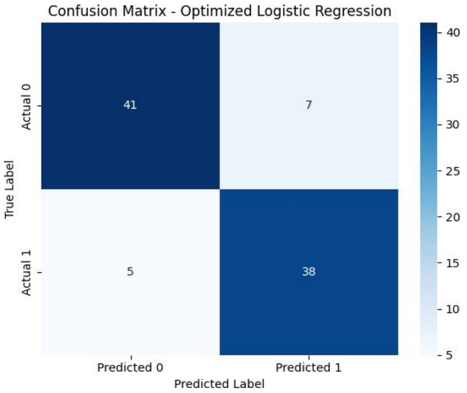
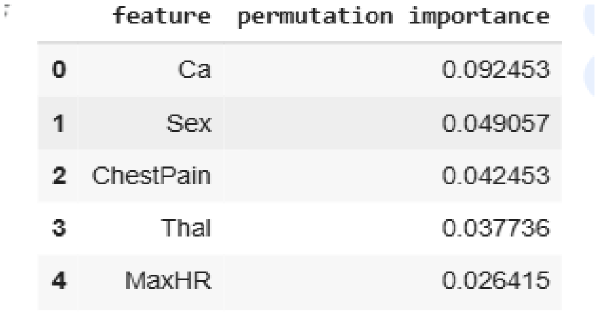


Fig6:Confusion matrixforoptimized logisticregression withWOA



	feature	permutation importance
0	Ca	0.092453
1	Sex	0.049057
2	ChestPain	0.042453
3	Thal	0.037736
4	MaxHR	0.026415

Fig:7. Permutationimportanceof features

Fig (7) displays the permutation importance of features in predicting cardiovascular disease using a machine learning model. Ca is the most influential feature with a permutation importance of 0.092453, indicating its strong impact on the model's predictions. This suggests that ca play a crucial role in diagnosing heart disease.

VI. CONCLUSION

In this pioneering research ,we have developed an optimized machine learning framework for cardiovascular illnesses prediction, combining advanced feature selection techniques with metaheuristic optimization to enhance accuracy and interpretability. Using SHAP analysis on Healthcare stroke dataset, we identified age, BMI,andaverageglucoselevelasthemostinfluentialfactorsinstrokeprediction,withageexertingthehighest impact.Additionalcontributors,suchasgender,residencetype,smokingstatus,andworktype,werealsofound to influence stroke risk, albeit to a lesser extent.To further validate the significance of clinical indicators in heart disease prediction, we employed permutation-based feature importance on Cleaveland heart disease dataset, revealing that ca, sex, chest pain, thalassemia and max heartrate plays a critical role in assessing cardiovasculararrisk. Additionally, theapplicationoftheWOAallowed usto fine-tuneourmodels,optimizing hyperparameters and significantly improving predictive performance. The optimized framework achieved a finalaccuracyof93%,withXGBoostdeliveringthebestresultsfortheHealthcarestrokedataset.Similarly,for the Cleveland heart disease dataset, the optimized Logistic regression model attained an accuracy of 87%, demonstrating its effectiveness in predicting cardiovascular risk. By integratingSHAP-driven insights, permutation-based feature selection, and WOA-driven optimization, we have created a powerful and interpretable machine learning framework. This approach not only achieves high accuracy but also enhances clinical decision-making by offering clear explanations for risk factors. Our framework serves as a valuable tool for healthcare professionals, enabling early diagnosis, personalized risk assessment, and proactive interventionstrategiestomitigateheartdiseaserisks.Futureenhancementsincludeextendingtheframeworkto otherdiseasessuchasdiabetes,cancer,Alzheimer's,andchronickidneydiseasetovalidateitsgeneralizability, andtestingitshybridapproachacrossdomainslikefinance,energy,andenvironmentalmodelingtoexploreits adaptability on large-scale datasets beyond healthcare

REFERENCES

[1] Hashi, E. K., & Md. Shahid Uz Zaman. (2020). Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. Journal of Applied Science &Amp; Engineering, 7(2), 631– 647. Process

[2] Dev,S.,Wang,H.,Nwosu,C. S.,Jain,N.,Veeravalli, B., & John,D.(2022). Apredictive analytics approach for stroke prediction using machine learning and neural networks. Healthcare Analytics, 2, 100088.

[3] Bhatt, C.M., Patel, P., Ghetia,T.,& Mazzeo, P.L. (2023).Effective Heart Disease Prediction Using Machine Techniques. Algorithms, 16, 88. Learning

[4] Karthick, K., Aruna, S.K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A.R. (2022). Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. Computational and Mathematical Methods in Medicine, 2022.

[5] Sunge, A.S., Amali, Zy, A.T., Pramudito, D.K., Badruzzaman, A., & Purwanto (2024). The model interpretability on SHAPand comparison classification selection feature for heart disease Computer Science. prediction. Procedia

[6] Nadimi-Shahraki, M., Zamani, H., Asghari Varzaneh, Z., & Mirjalili, S. (2023). A Systematic Review of the Whale Optimization Algorithm: Foundation, Theoretical Improvements, and Hybridizations. Archives of Computational Methods in Engineering, 1 - 47.

[7] Sayegh, H.R., Dong, W., & Al-madani, A.M. (2024). Enhanced Intrusion Detection with LSTM-Based Model, Feature Selection, and SMOTE for Imbalanced Data. Applied Sciences.

[8] Tian, T., Liang, Z., Wei, Y., Luo, Q., & Zhou, Y. (2024). Hybrid Whale Optimization with a Firefly Algorithm for Function Optimization and Mobile Robot Path Planning. Biomimetics, 9.

[9] Rezk,N.G.,Alshathri,S.I.,Sayed,A.,ElDinHemdan,E.,&El-Behery,H.(2024).XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach. Bioengineering, 11.

[10] Nagavelli,U.,Samanta, D.,& Chakraborty,P.(2022).Machine Learning Technology-Based Heart Disease Detection Models. Journal of Healthcare Engineering, 2022.

[11] Ahmed, H., Younis, E. M., Hendawi, A., &Ali,A.A. (2020). Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generation Computer Systems, Vol.111, 714 722.