CHAPTER 3

RESEARCH METHODOLOGY

3.1 OVERVIEW

In the case of heart disorders, in particular, early disease prediction plays a crucial role in healthcare. One of the most intriguing and difficult issues is the development of ML models to effectively predict the risk of heart illness. In order to better anticipate the threats of cardiovascular illness, researchers are aiming to improve the accuracy and execution speed of existing ML models. Choosing optimal features is becoming more and more necessary as dataset dimensions increase in predictive modeling. Existing feature selection approaches still face difficulties with datasets that have complicated feature relationships and significant degrees of redundancy. Ensemble classifiers have been shown to outperform individual classifiers in classification tasks. Ensemble methods including majority voting, majority averaging, majority based, and a meta-classifier are now available. However, the best classifier ensemble may be dynamically formed togetherwith the best representative subset of classifiers, rather than comprising all classifiers. Thelimitations of the available feature selection methods and classifiers are effectively identified and new methods are proposed in this study.

3.2 PROPOSED METHODOLOGY

3.2.1 Selection of highly relevant features for heart disease prediction

There is no linear correlation between the number of features and model performance. Until a certain threshold value, the ML model would exhibit high performance, but on adding more dimensions or features, the performance of the model would decline. This "Curse of Dimensionality" due to iterating over highly correlated predictor features and also the use of insignificant features toward predicting the target variable would decrease the performance of an ML model and consume more time. So, selecting highly contributing features toward the prediction of the target variable is extremely important.

3.2.1.1 Feature Selection using Importance Scores of features by Gradient Boosting Classifiers

Feature Importance (FI) enables us to comprehend the importance of each input feature while making predictions of the target variable. In this study, the importance scores of the attributes by gradient boosting classifiers, XGBoost and CatB, were obtainedon heart datasets. Applying the forward selection method yields subsets of features using the FI score as the threshold. CatB , Hard Maj. Vote Ensemble (comprising CatB, GNB, LR, XGBoost, RF), and XGBoost, are used for classification. The optimal feature subset which yields the highest accuracy and its possible range are determined.

3.2.1.2 Feature Selection by ModifiedBoostARoota and heart disease prediction using classification models

To pick the important characteristics for the prediction of heart disease with reduced computational cost, this study offers a new wrapper feature selection approach called ModifiedBoostARoota (MBAR). The MBAR algorithm is a modified version of the 'BoostARoota (BAR)' algorithm (Chasedehan, 2018). MBAR differs from BAR by utilizing a different base model and using different feature elimination process in the algorithm. A comparison of MBAR and BAR algorithms is done by experimenting using both CatB and XGBoost as base models in each. Following the feature selection process, CatB and XGB are modeled on the selected set of features and experimental outcomes of MBAR are compared with contemporary methods of choosing features in terms of prediction performance on a number of heart illness datasets.

3.2.2 A Novel Super Learner Ensemble Learning Model in Heart Disease Prediction

An individual ML classifier is vulnerable to noise, outliers, and dataset biases. Poor generalization may result from inaccurate predictions or sensitivity to outliers in the training data. Ensemble classifiers are generally more robust due to their ability to averageor vote on multiple predictions. They can mitigate the impact of individual model errors oroutliers, making them more resilient to noise and improving overall stability. Classifiers that have had their hyper parameters fine-tuned produce good results on some datasets but not others. To improve upon the performance of individual base classifiers, it is possibleto combine them into a single ensemble model. Combining models to increase prediction

accuracy has been the subject of numerous theoretical and empirical investigations. Ensemble models integrate the forecasts of multiple standalone models.

In this study, we choose as our starting points a number of ML classifiers, including the Support Vector Machine (SVM), LR, XGBoost (XGB), CatB, RF, KNN,DT, and Gaussian NB (GNB). These classifiers are trained on the heart datasets via iterative rounds of stratified k-fold cross-validation. The ensemble of classifiers provides more trustworthy results than any individual classifier. In contrast to more sophisticated statistical models, however, ensemble models cannot reliably draw valid conclusions about the predictors of interest. This problem is solved by the proposed Super Learner Ensemble Model (SLEM) which employs a meta learner to integrate predictions from multiple base models. The suggested ensemble model, SLEM, benefits from a new mix of basic models that enhances its performance.

3.2.3 An Optimized Super Learner Ensemble Model Using Whale Optimization Algorithm

The effectiveness of any one ML classifier may be capped by the problem's complexity and the model's inability to capture all the underlying patterns. When compared to individual classifiers, ensemble classifiers may get better results. The ideal combination of multiple models in an ensemble model can reduce bias, variance and overfitting, leading to better generalization and higher prediction accuracy.

In this section of the research, the Whale Optimization Algorithm (WOA) isutilized to select appropriate classifiers for the ensemble, hence enhancing the ensemble model. By taking into account both the accuracy of the combined classifiers and their pairwise diversity, the accuracy-diversity-based pruning strategy is introduced in the context of ensemble pruning.

In the proposed work, Initially, based on the diversity of the learning models, SVM, LR, GNB, RF, KNN, DT, MVE, XGB and CatB are chosen as base classifiers. WOA algorithm uses "whales" to indicate the existence or absence of key foundational classifiers. The WOA outputs a whale with good fitness value. This is repeated fordifferent sets of iterations. Then, the diversity measure value is calculated for each whale, and the most diverse whale is chosen. Super Learner Ensemble Model (SLEM) finalizes

its base models with the specified whale (basic classifiers). The SLEM ensemble classifier takes as input the combined predictions of all the basis classifiers, and the meta learner, LR, improves its performance over time by more effectively combining the results of eachof the foundation models. OSLEM is the name of the proposed study.

3.3 METHODOLOGY FLOW

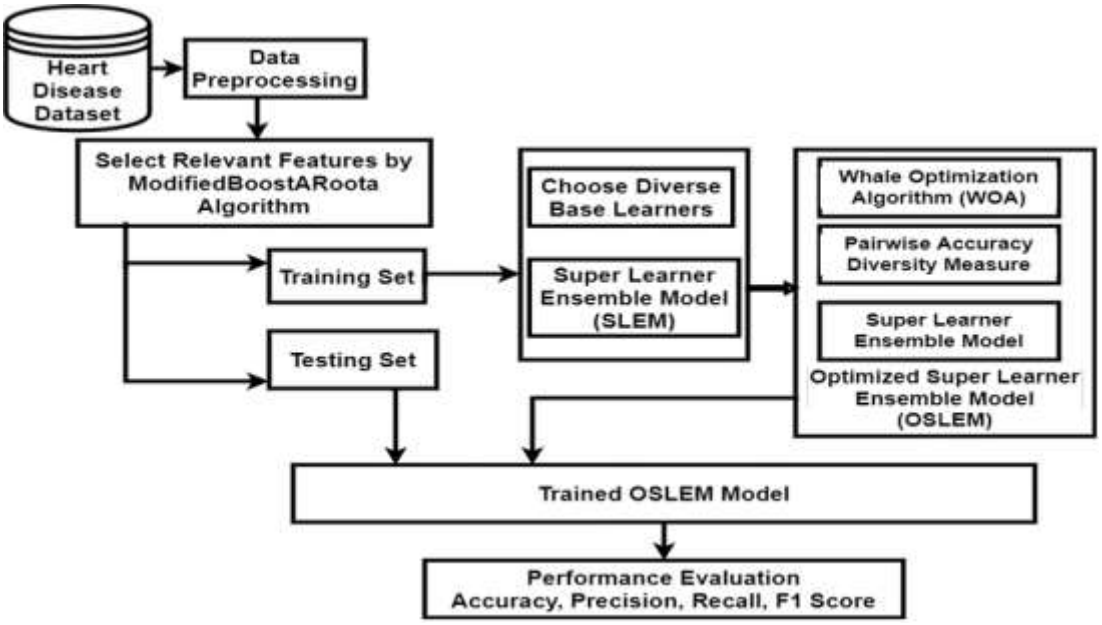

Figure 3.1 Block Diagram of Proposed Research

3.4 DATASET DESCRIPTION

To investigate the proposed model's performance, seven Heart Disease (HD) datasets are used and described as follows:

1. **Cleveland Heart Dataset:** This dataset is hosted at UC Irvine ML Repository. There are total of 303 records with 13 predictor attributes and one target attribute. Attributes are classified as either real (1,4,5,8,10,12) or ordered (11), binary (2,6,9) or nominal (7,3,13). Age, sex, chest pain classification (4 values), resting blood pressure, serum cholesterol in milligrams per deciliter, fasting glucose > 120 milligrams per deciliter, resting ECG findings (0, 1, 2), maximal heart rate, exercise-induced angina, and resting ECG results (0, 1, 2) are all variables. Three primary vessels (on fluoroscopy; normal range: 0-3), a steep ST segment at the peak of activity, and a normal thal: normal range: 6-7. The result or expected

variable 'num' takes on the values 1, 2, 3, and 4 if cardiac disease is present, and 0 otherwise. In this research, the target variable has all heart illness instances grouped under class 1 and no heart disease instances are grouped under class 0. The attributes are shown in Table 3.1.

Table 3.1 Features in Cleveland Heart Dataset

S. No.	Names of Features	Range of Values
1.	Age of the Patient (age)	29 to 77
2.	Gender of the Patient (sex)	0 or 1
3.	Chest Pain type (cp)	1 to 4
4.	Blood pressure-Resting (trestbps)	94 to 200
5.	Cholesterol-Serum (chol)	126 to 564
6.	Blood Sugar-Fasting (fbs)	0 or 1
7.	ECG-Resting (restecg)	0 to 2
8.	Heart rate-Max. (thalach)	71 to 2 02
9.	Exercise induced angina (exang)	0 or 1
10.	ST depression induced by exercise (oldpeak)	0.0 to 6.2
11.	Slope of the peak exercise ST segment (slope)	1 to 3
12.	Number of major vessels colored (ca)	0 to 3
13.	Defect type (thal)	3,6 or 7
14.	Target attribute (Num)	0 or 1

2. **Statlog Heart Dataset:** This data is hosted at the ML Repository of UC Irvine. There are 270 records total, with 13 unique qualities (one of which is dependent). There are two possible values for the dependent variable: 1 if the disease is present, and 2 if it is not. Attributes are offered as either real (1,4,5,8,10,12) or ordered (11), binary (2,6,9) or nominal (7,3,13) types. Attributes include: age, sex, kind of chest discomfort (4 values), blood pressure at rest, serum cholesterol in milligrams per decilitre, fasting blood sugar > 120 milligrams per decilitre, resting electrocardiographic results (0, 1, 2), and maximal heart rate. EIA, oldpeak = ST depression during exercise in comparison to rest, main vascular coloration as

evaluated by fluoroscopy (0-3), ST segment slope at the peak of a workout, thal: 3 for typical, 6 for a permanent flaw, and 7 for a correctable flaw. The target variable has all no heart disease instances are grouped under class 1 and all heart disease instances grouped under class 2. Table 3.2 lists the characteristics.

Table 3.2 Features in Statlog Heart Dataset

S. No.	Names of Features	Range of Values
1.	Age of the Patient (age)	29 to 77
2.	Gender of the Patient (sex)	0 or 1
3.	Chest Pain type (cp)	1 to 4
4.	Blood pressure-Resting (restbp)	94 to 200
5.	Cholesterol-Serum (chol)	126 to 564
6.	Blood Sugar-Fasting (fbs)	0 to 1
7.	ECG-Resting (restecg)	0 to 2
8.	Heart rate-Max.(maxheartrate)	71 to 202
9.	Exercise Induced Angina **(**angina)	0 to 1
10.	ST depression induced by exercise (oldpeak)	0.0 to 6.2
11.	Slope of the peak exercise ST segment(slope)	1 to 3
12.	Number of major vessels colored (coloredvessels)	0 to 3
13.	Defect Type (thal)	3,6 or 7
14.	Target Attribute (disease)	1 or 2

3. **South African (SA) Heart Dataset:** This dataset is hosted at Harvard Dataverse. Class, a dependent variable, is one of 10 aspects, of which 9 are independent clinical qualities. There are 462 medical observations. For the medical observations, the class variable is either 1 (indicating no coronary heart disease event was found) or 0 (indicating that an incident was observed). Men from a high-risk area of South Africa for cardiovascular disease make up this sample. Table 3.3 displays the features retrieved while keeping an eye on each high-risk patient inthe sample.

Table 3.3 Attributes in SA Heart Dataset

S. No.	Names of Features	Range
1.	**adiposity** –Percentage of body fat	[6.74, 42.49]
2.	**obesity** -weight-to-height ratio or body mass index)	[14.7, 46.58]
3.	**sbp** -Systolic blood pressure	[101, 218]
4.	**tobacco** -accumulative tobacco in the body	[0.0, 31.2]
5.	**ldl** -Low density cholesterol	[0.98, 15.33]
6.	**famhist** -family history of heart disease	[0, 1]
7.	**typea** -Type A behaviour and personality	[13, 78]
8.	**alcohol** -current alcohol consumption	[0,147.19]
9.	**age**	[15, 64]
10.	**class**	{-1,1}

4. **Cardiovascular Disease Dataset:** This dataset is hosted at Mendeley data. This dataset consists of 1000 instances with 12 predictor variables and 1 target variable. The target variable has all heart illness instances grouped under class 1 and no heart illness instances are grouped under class 0. The variables are shown in Table 3.4.

Table 3.4 Features in Cardiovascular Disease Dataset

S. No.	Names of Features	Range of Values
1.	Age of the Patient (age)	[20, 80]
2.	Gender of the Patient (sex)	[0,1]
3.	Chestpain (Chest Pain type)	[0, 3]
4.	restingBP (Resting blood pressure)	[94, 200]
5.	serumcholestrol (Serum cholesterol)	[0, 602]
6.	fastingbloodsugar (Fasting Blood Sugar) (≤ or >120)	[0, 1]
7.	restingrelectro (ECG at rest)	[0, 2]
8.	maxheartrate (Maximum heart rate)	[71, 202]
9.	exerciseangia (Exercise induced angina)	[0, 1]
10.	Oldpeak (ST depression induced by exercise)	[0.0, 6.2]
11.	Slope (Slope of the peak exercise ST segment)	[1, 3]
12.	noofmajorvessels (Number of major vessels colored)	[0, 3]
13.	target -Predicted attribute	{0,1}

5. **Arrhythmia Heart Dataset:** ML Repository of University of Irvine hosts this dataset, which consists of ECG signals data with 279 attributes and 452 instances. Among the attributes, 206 contained linear values, and the rest are nominal. The instances of the dataset belonged to sixteen groups or classes. Class 1 referred to normal beats. Class 2 to Class 15 referred to different types of Arrhythmias. Unclassified beats were grouped as in Class 16. There are 245 instances of normal types, and 207 instances of the abnormal types. In this research work, theseinstances are grouped into two classes: i) normal and ii) arrhythmia.
6. **Z-Alizadeh Sani Heart Dataset:** It is available in the UCI Machine Learning Repository. The dataset consists of 55 predictor features and 1 predicted feature, related to CAD and instances are 303. Features 1 to 17 are demographic features, features 18-31 are symptoms and examinations category, features 32-38 are from ECG data, 39-54 are laboratory and Echo related. Feature 55 is for valular heart diease and feature 56, 'Cath', is the target variable. If the diameter narrowing is lesser than 50% a patient is considered as normal (value 0), else the patient has coronary artery disease (value 1). The list of variables are in annexure-I.
7. **Cardiac Biomarkers Dataset:** This is a real world dataset collected from Specialist hospital, Bangalore, containing the data of 192 patients who presented chest pain and had undergone lab tests in order to confirm whether they had Acute Myocardial Infraction (AMI) or not. The personal details of the patients were not disclosed. The data collected consists of independent variables: Age, Gender,CKMB (Creatine Kinase Myocardial Band), Myoglobin, Troponin-I, BNP (Brain Natriuretic Peptide), D-Dimer, ACS_types and the dependent variable is Disease with class 0- no AMI, class 1- AMI, class 2- heart problems. There are 61 patients without AMI, 3 patients in critical heart failure, 51 patients with AMI and heart muscle damage, 42 patients with only cardiac muscle damage, 18 patients diagnosed with AMI within 8 hours, 7 diagnosed within 8 hours and again after 24 hours, 1 patient diagnosed with AMI after 24 hours, 5 patients with muscle/skeletal damage, and 4 patients with higher-level blood clots in this dataset. To summarize, there were 61 patients who did not have an AMI and so belonged to illness class 0, 80 patients who did have an AMI and thus belonged to disease class 1, and 51 patients who had cardiac muscle injury and clots and thus

belonged to disease class 2. In classes 0 and 2, the proportion of male to female patients is nearly equal, but in class 1, which includes AMI patients, the male to female patient ratio is higher. A 0 is entered for males and a 1 for females in the gender field. Each of these biomarkers has a normal range (in the absence of AMI), D-Dimer: 0 to 400, BNP: 0.0 to 100, CKMB: 0.0 to 4.3, Troponin: 0.0 to 0.4, and Myoglobin: 20 to 80, where the measuring scale is nanograms per millilitre. The Attributes are shown in Table 3.5.

Table 3.5 Features in Cardiac Biomarkers Dataset

S. No.	Names of Features	Range of Values
1.	Age of the Patient (age)	21 to 90
2.	Gender of the Patient (gender)	0 or 1
3.	CKMB- Creatine Kinase Myocardial Band	[1, 80]
4.	Myoglobin	[8.6,500]
5.	Troponin-I	[0.01,30]
6.	BNP - Brain Natriuretic Peptide	[4,4900]
7.	D-Dimer	[100,5000]
8.	ACS_types -Heart disease types	[0,10]
9.	Disease- Target Variable	{0,1,2}

3.5 PERFORMANCE METRICS

The effectiveness of the models developed is evaluated using a variety of measures. Brief descriptions of each are provided below:

3.5.1 Accuracy

The accuracy rate measures how many samples or labels were correctly predicted to have heart disease. It is determined as

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + False\ Positive\ (FP) + False\ Negative\ (FN)} \quad (3.1)$$

- TP measures an outcome for which the classifier classifies Heart Disease labels precisely as Heart Disease.

- FP quantifies a result of classifier which incorrectly classifies instances of Heart Disease labels as non-Heart Disease labels.
- FN measures an outcome in which the classifier incorrectly classifies non-Heart Disease labels as Heart Disease labels.
- TN measures a result for which the classifier classifies non-Heart Disease labels precisely as non- Heart Disease.

3.5.2 Precision

It provides a metric for the proportion of projected positive observations that really turn out to be positive (having heart disease).

$$Precision = \frac{TP}{TP+FP} \tag{3.2}$$

3.5.3 Recall

Recall measures how many positive (heart illness) observations the classifier correctly identifies out of the total number of positive (heart disease) observations.

$$Recall = \frac{TP}{TP+FN} \tag{3.3}$$

3.5.4 F1-Score

When the class is imbalanced, the ideal measure is F1-Score. F1-Score is the harmonic mean of recall and precision. Maximising the F1 score requires concurrently maximising both recall and precision since it combines them using their harmonic means.

$$F\ Score = \frac{2 \times precision \times Recall}{precision\ + Recall} \tag{3.4}$$

3.6 CHAPTER SUMMARY

This section presents the overall approach taken in this study. The next chapters elaborate on the methodology behind the proposed methods. In addition, the experimental datasets are discussed. The performance measures employed in this proposed work are also described.