CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

This chapter presents a comprehensive overview of previous research on feature selection methods, Machine Learning methods, Optimization of Machine Learning methods and Ensemble classifiers for disease prediction. In addition, the limitations of the approaches are also tabulated for providing an idea to propose new methods to solve these limitations.

2.2 RELATED WORKS

2.2.1 Review of Feature Selection Methods for Disease Classification

Kavitha et al., (2016) proposed the dimensionality reduction technique to compress the high dimensional data to lower dimensional data with some constraints. A framework for the easy prediction of heart disease was created by using the Principal Component Analysis (PCA) to extract the features and a mathematical model was computed to select the relevant features using the relevant constraint. Principal components are linear combinations of the features from the original data, but they are not as easy to interpret for large-size data. This work helps in improving the efficiency, accuracy, and speed of the process. But, it has low interpretability of principal components.

A unique hybrid approach for detecting and verifying CAD instances was described by Verma et al. (2016). Using a PSO search strategy for data preprocessing with a correlation-based feature subset selection to predict coronary heart disease is a hybrid approach. PSO can reduce the dimensionality of a data set, making the system simpler. It can be a useful supplementary tool in clinical settings because it allows for reliable and objective diagnosis. Compared to CFS and PSO, the proposed hybrid method performs better in picking features for predicting CAD, and the experimental findings also demonstrate that the model can be implemented with minimal amounts of clinical data. The goal of PSO is to lessen the number of dimensions in a dataset, which in turn simplifies the system.

The Ranker algorithm was proposed by Sabab et al. (2016) to prioritize and remove irrelevant features. Any classification method can be vastly enhanced by employing the right attribute selection approach. Based on the attributes present in the dataset, three classifiers have been selected. They are the NB, SVM, and DT methods. The algorithm used in this feature selection method is sophisticated and requires training before it can be used effectively.

To address the lack of population diversity in the traditional PSO algorithm, (Vijayashree et al., 2018) developed a fine tuning technique. The fine-tuning procedure selects the optimal weight for updating the particle velocities and locations. Several well-known feature-selection approaches, such as Info gain, Chi-squared (CS), one-attribute based, Consistency subset, Relief, CFS, Filtered subset, Filtered attribute, Gain ratio, and the PSO algorithm, are compared to and contrasted with the proposed PSO-SVM. Nonetheless, it has a hard time determining an appropriate weight to update the velocities and positions of the particles.

To enhance illness classification rates and try out different disease patterns, Naganjaneyulu et al. (2018) suggested a unique feature selection based classification approach. Weighted feed forward neural network was used to enhance classifiers like SVM and DT in a proposed model. One major negative is the increased number of tuning settings.

A novel wrapper feature selection technique was proposed by Al-Tashi et al. (2019) to determine the optimal feature subset for CAD diagnosis. The suggested method has two key phases: selecting features and classifying them. The first stage was to use Grey Wolf Optimization (GWO) to determine which aspects of the disease detection dataset would prove most useful. Second, the GWO fitness function is tested using a Support Vector Machine (SVM) analysis. The effectiveness of the proposed technique is evaluated using the Cleveland Heart disease dataset, which is open to the public. The experimental results demonstrated that the proposed method outperformed the state-of- the-art methods in terms of accuracy, sensitivity, and specificity. The major negative was the time and effort required to compute.

Gokulnath et al. (2019) proposed using a SVM as the basis for an optimization function. The more important variables for cardiac disease prediction are selected using

this objective function in a GA. The experimental outcomes of the Gawith SVM are contrasted with those of other existing feature selection strategies, such as Relief, CFS, Filtered subset, Info gain, Consistency subset, Chi squared, one attribute based, Filtered attribute, Gain ratio, and GA. The same genetic algorithm (GA) with Binary Particle Swarm Optimization (BPSO) is used in both tasks. In addition, many other classification and regression methods are used to mine the healthcare dataset for previously undiscovered nuggets of gold. Yet, the computational burden is heavy.

Jeyaraman et al. (2019) provide a particle bee optimized associative memoryneural network and an artificial gravitational cuckoo search algorithm to handle the peculiarities of the original heart disease categorization system. At first, information is culled from the Heart Disease Data Set-UCI database. Because of the size and complexity of the data obtained, the effectiveness of the system for identifying cardiac disease is diminished. So, characteristics are becoming less complex as predicted by the gravitational cuckoo search algorithm's behavior. The defined associative memory classifier is applied to the chosen features. The drawback is to need more time to practice the system.

By developing a Correlation-based Feature Selection Subset Evaluator as Feature evaluator and employing Greedy Stepwise as a search technique for feature selection, Singh et al. (2020) improved their ability to predict liver disease. The suggested work usesa dataset of liver disorders suffered by patients to test the efficacy of various classifiers, including the NB Classifier, LR, and Sequential Minimal Optimization (SMO). LR classifier with feature selection approaches produced the best results, and the use of feature selection technique lowered the execution time of various classifiers. But, it increasing overfitting risk when the number of observations is insufficient.

In light of the problems caused by feature attribute dimension disasters in diagnosis decision support systems, Huang et al. (2020) suggested an association rule- based algorithm for selecting features. The experimental results reported here show that the RBSBagging method described in this work is both accurate and effective. Theproposed association rule-based feature selection method ARFS minimizes the size of feature subsets without sacrificing classification precision. However, this reduces a

model's explainability. The problem with RBSBagging is that it reduces a model's readability.

Thyroid classification procedures can benefit from applying the Group GWO computation, as indicated by Shankar et al. (2020). The data is classified as hypothyroid, hyperthyroid, or normal using a classifier model trained using the support vector machine (SKSM). The experimental outcome results in better metrics across the board, including accuracy, sensitivity, and specificity. The results of using this method to spot thyroid disease will be promising. High computational time is a limitation of this effort.

Garate et al. (2020) suggested combining PCA with CS tests to boost the accuracy of ML model predictions. Classifier's mission was to determine if a person has heart disease. Cholesterol, maximum heart rate, chest discomfort, features linked with ST depression, and cardiac vessels were all retrieved using CS because of their anatomical and physiological significance. This method has broad potential for usage in both medical diagnostics and real-world contexts because to its ability to process enormous data setsand identify the root causes of a variety of conditions. The primary issue is that the samplesize is insufficient to draw any firm conclusions about the prevalence of heart disease in the population as a whole.

A Recursive feature elimination and GA feature selection strategy was proposed by Divya et al. (2021) to enhance the classifier's sensitivity and specificity in distinguishing MCI from AD. Wrapper-based feature selection approaches like RFE and GA are used to discover the optimal set of features to employ. Classification performance is evaluated using the Mini-Mental State Examination (MMSE) score and the best volumetric segmentation features. But it can be difficult to optimize and can be time- consuming to run.

A classification system for Parkinson's disease (Pd) based on vocal features obtained from speech recordings was presented by Gunduz et al. (2021) (using a proposed hybrid DR approach to extract robust features from voice recordings). Both filter-based feature selection models and variational Autoencoders (VAE) saw their strengths combined in the suggested approach. Feature extractor VAE was selected due to itscapacity to preserve the regular latent space properties throughout the feature generation process, whereas filter-based methods Relief and Fisher Score were selected because to

their efficacy in handling noisy data. For this purpose, deep feature representations were created and used to train a multi-kernel SVM classifier.

Riajuliislam et al. (2021) used Recursive Feature Selection (RFE), and compared it with Univariate Feature Selection (UFS) and Principal Component Analysis (PCA) along with five different classification algorithms namely, Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Naive Bayes (NB) to predict the Thyroid disease in early stage. The RFE along with all the classifiers gave the same accuracy and higher than other techniques. The model was experimented only on a dataset taken from Bangladesh clinic and not on other datasets.

To pick the right features for heart disease detection, Rajadevi et al. (2021) developed the Black Hole Optimization (BHO) algorithm as a suitable feature selection method. When compared to other classification models like KNN, DT, RF, and SVM, XGBoost+ BHO showed excellent performance.

Mandal et al. (2021) provide a tri-stage wrapper-filter based feature selection architecture for the purpose of disease detection from medical reports. Incorrect classification due to the removal of any very relevant trait is especially undesirable in the context of medical data. To pick each feature with the maximum accuracy independent of the filter method or classification algorithm employed, Phase 1 of the method combines three classification algorithms (KNN, SVM, and NB) with four filter methods (Mutual Information (MI), ReliefF (RFF), Chi Square (CS), and Xvariance (XV)). Next, the union set is sent through the XGBoost classification algorithm, and k top features are chosen based on their accuracy. But, higher risk of over-fitting than deterministic algorithms.

For classification, a deep convolutional neural network was suggested (Nagarajan et al., 2022), while a Genetic-Based Crowd Search Algorithm (GCSA) was recommended to pick features. When compared to existing feature selection strategies like DT, support vector machine, and artificial crow colony, the suggested technique achieves very high accuracy. This task tends to become stuck in a local optimum and has a slow convergence rate.

In order to create a highly accurate CKD diagnosis system, Hossain et al. (2022) analyze numerous feature optimization approaches in addition to a max voting ensemble

model. Feature selection is the process of identifying the most important features from a pool of candidates and eliminating the rest. In this research, we employ two state-of-the- art feature selection techniques: minimum redundancy maximum relevance (mRMR) and recursive feature elimination (RFE). In this investigation, an ensemble classifier evaluates distinct optimized feature sets generated via several feature optimization strategies. In this study, the max voting ensemble model is built by combining five widely-used classification models: LR, RF, SVM, KNN, and Xtreme gradient boosting (XGB). The primary issue is the increased need for MI computations between any given feature and itsclass label or any pair of input features.

To categorize AD using volumetric and statistical data from brain MRIs, Keles et al. (2022) presented a binary variation of the artificial bee colony algorithm (BABC). The MRIs were provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI). The gathered MRIs were uploaded to the online portal volBrain, where volumetric and statistical data were obtained. After that, we used BPSO, BGWO, and BDE to evaluate their differences and similarities. The feature selection process is compared using the classifiers KNN, RF, and SVM. But, it suffers from improper exploitation in solving complicated problems.

Table 2.1 Comparison of Feature selection methods for disease classification

Author & year	Method	Merits	Demerits
Kavitha et.al, 2016	principal component analysis (PCA) framework	Simple to calculate. Since PCA is based on linear algebra, it can be quickly and easily solved by computers.	Principal components have a low level of interpretability. Principal components, linear combinations of features in the original data, are not as intuitive to examine.
Verma et.al, 2016	correlation-based feature subset selection with PSO search method	It may significantly improve the accuracy of identifying people with CAD.	In order to reduce the system's complexity, PSO uses dimensionality reduction.

Author & year	Method	Merits	Demerits
Sabab et.al, 2016	Ranker algorithm	The algorithm has a high degree of accuracy.	It is complex and require training to use
Vijayashree et.al, 2018	PSO algorithm, SVM	The dataset is cleaned of any superfluous or duplicate information as well.	It has trouble picking the right weight to update the particle velocities and positions.
Naganjaneyulu et.al, 2018	Feed Forward Neural network	It can handle and process nonlinear data easily compared to perceptrons and sigmoid neurons	More parameters to optimize.
Al-Tashi et.al, 2019	GWO, SVM	It Provides flexibility with the concept of kernels	Computationally expensive
Gokulnath et al, 2019	GA, SVM	It is good for noisy environment	It is Computationally high.
Jeyaraman et al, 2019	artificial gravitational cuckoo search algorithm	System performance in terms of available metrics is analyzed.	Need more time to practice the system
Singh et.al, 2020	Correlation-based Feature Selection Subset Evaluator	It helps doctors make more informed decisions in the clinic.	Increasing overfitting risk when the number of observations is insufficient.
Huang et.al, 2020	RBSBagging algorithm	giving a group of average students a fighting chance against a single top student	It reduces a model's explanatory power.
Shanker et al, 2020	Group Grey Wolf Optimization (GGWO)	The performance of accuracy was improved.	It implicitly assumes that all the attributes are mutually independent.

Author & year	Method	Merits	Demerits
Garate et.al, 2020	chi-square (CHI) with PCA	It can be used in various medical diagnosis and real-world applications to analyze large amounts of data and spot potential health problems.	Due to the small sample size, it is challenging to generalize these results on cardiovascular disease.
Divya et.al, 2021	RFE and GA	Improved performance in terms of predictive accuracy	It can be difficult to optimize and can be time-consuming to run
Gunduz et.al, 2021	Hybrid Feature Reduction Mechanism	Models that are more flexible and robust against high dimensional data.	Low interpretability of the model
Riajuliislam et.al, 2021	Recursive Feature Selection(RFE), Univariate Feature Selection(UFS) and Principal Component Analysis(PCA)	Speeding up the training process and decreasing the amount of data required go hand in hand.	Selecting a minimal collection of features that adequately represents the topic being described.
Rajadevi et.al, 2021	Black Hole Optimization (BHO) Feature Selection algorithm with XGBoost classifier	Improved Performance Metrics	Only one dataset was used
Mandal et.al, 2021	wrapper-filter based feature selection methods	Improves Accuracy.	Higher risk of over-fitting than deterministic algorithms
Nagarajan et.al, 2022	Genetic-Based Crow Search Algorithm	The method achieves better accuracy than	Convergence is slow, and it's simple

Author & year	Method	Merits	Demerits
		other FS strategies, including the Decision tree, the support vector machine, and the artificial crow colony.	to get stuck in a "local optimum."
Hossain et.al, 2022	Minimum redundancy maximum relevance (mRMR) and Recursive feature elimination (RFE)	Remove unnecessary information without losing data that the model needs.	The number of MI computations performed between a feature and its class label, or between a feature and another input feature, increases.
Keles et.al, 2022	artificial bee colony algorithm	Outperforms all traditional methods considered in the study	study is based on a specific dataset, the results may not be generalizable

2.2.2 Review of Machine Learning Methods for Disease Classification

For the purpose of predicting CAD, Nassif et al. (2018) suggested a ML classification. Accurate categorization requires thoughtful choice of methods and features. SVM, NB, and KNN classifiers were trained using 10-fold cross-validation, with the features as inputs. Three feature selection strategies and three classification algorithms (SVM, NB, and KNN) are compared and contrasted in this work. The Naive Bayes classifier outperforms SVM and KNN on the Cleveland dataset, according to the results. One potential flaw in this study is that the categorization algorithms take into account qualities with a large number of possible values.

The Hybrid RF with Linear Model (HRFLM) was developed by Mohan et al. (2019). Improving the accuracy with which heart illness may be anticipated is the primary focus of this investigation. The suggested HRFLM method combines aspects of both the RF andthe LM into a single powerful tool. HRFLM's predictive abilities for cardiovascular

disease were shown to be high. Different combinations of ML methods can be used to improve prediction methods in the future of this study. The proposed hybrid strategy outperforms gold-standard methods in predicting cardiovascular disease, as shown by our experiments. Training is a time-consuming process.

Khourdifi et al. (2019) employed the Fast Correlation-Based Feature Selection (FCBF) method to reduce irrelevant features in order to improve the precision of heart disease classification. The suggested hybrid method is used to a heart disease dataset, and the findings show that it is effective and robust at processing different forms of data for heart disease classification. PSO is used to optimize a system, and then the PSO result is used as a starting point for ACO. However, when the number of variables is huge, the computation time becomes prohibitively long.

According to Reddy et al. (2019), a combined dataset was used to evaluate the CARET package of the R tool for cardiovascular disease prediction, including its classification and feature selection methods. The stochastic gradient boosting approach and the recursive Feature Elimination (RFE) method were used to narrow down the characteristics to chest pain, ca, and Thal. However, it is higher likelihood of overfitting, model harder to explain and costlier to maintain.

Tougui et al. (2020) examined the efficacy of six ML techniques for disease classification, including LR, support vector machine, k-nearest neighbors, AI, NB, and RF, with heart disease as an example. The diagnostic judgments made by doctors, physicians, and other medical professionals would benefit from the availability of such technologies. However, higher system performance can be achieved by locating the bias value in an economical and time-saving approach.

Using ML methods, Badriyah et al. (2020) classified stroke data from CT scan pictures. Before classifying images, the data must be processed and features extracted. Features such as contrast, dissimilarity, homogeneity, energy, correlation, and automated structure matching (ASM) are used in the extraction of features from the Gray-Level Co-Occurrence Matrix (GLCM). Classification methods are then evaluated, including K-Nearest Neighbors, NB, LR, DT, RF, MLP-NN, Deep Learning, and SVM. When compared to the other categorization algorithms used in this study, RF yielded the highest

accuracy. Classification accuracy using optimization's default parameter value has not been tested, though.

An ML-based technique for diagnosing heart disease was presented by Li et al. (2020). The experimental findings show that the proposed ANN and Fuzzy Logic features selection method selects features with higher effectiveness and greater classification accuracy than the conventional feature selection algorithms. Thallium scan, chest pain and exercise-induced angina are the most important symptoms to monitor, according to most feature selection algorithms. All FS algorithm results point to the fact that fasting blood sugar (FBS) is not an accurate predictor of cardiovascular disease. For HD detection, the ML-based CMIMSVM technique outperforms the deep neural network. The problem is that it takes more time to execute before a result is produced.

Using Genetic Algorithm and Convolutional Neural Networks (GA-CNN) and GA-SAE (Genetic Algorithm and Stacked Autoencoder) hybrid deep learning models, Kilicarslan et al. (2021) set out to create a decision support system to aid clinicians in the correct classification and diagnosis of anemia. The hybrid models provided here have had their hyperparameters tuned by a GA. These models analyze medical data for diagnosis, lowering the likelihood of mistakes due to human inexperience or exhaustion. Other intelligent optimization strategies are planned for use in the future to improve the success rate of obtaining hyperparameters of deep learning algorithms. It is highly complexity to work.

The synthetic minority over-sampling method proposed by Lee et al. (2021) and implemented with TomekLinks was used to correct the data imbalance. This research shown that gradient boosting with synthetic minority over-sampling is a good diagnostic strategy for binary classification and multi-classification schemes, with the ability to forecast a parasite state. It's unbalanced and lacks in depth, thus it can't replace the real thing.

A new model suggested by Gopal et al. (2021) aims to use ML to detect early-stage breast cancer. The correlation coefficient function and multi-filtering techniques were used for feature selection, which allowed for the efficient selection of characteristics. In order to classify cases of breast cancer (BC), the attribute evaluator used PCA based on

a ranking algorithm. Accuracy of the prediction approach was enhanced by using the MLP classifier. The demerits of this method is it has a limitation of specific number of dataset.

Correlation Based Feature Selection (CFS), Gain Ratio, Hidden Markov Model (HMM), Artificial Neural Network (ANN), SVM, and DT J48 are all analyzed in detail in the procedure provided by Archika et al. (2022). The proposed method fixes problems with both older and newer methods. The fundamental problem arises when there are many possible values for an attribute, which might cause overfitting.

Table 2.2 Comparison of ML Methods for Disease Classification

Author & year	Method	Merits	Demerits
Nassif et al, 2018	SVM, Naive Bayes and KNN	dependant on one another more linearly, and so impact the dependent variable similarly	Overfitting is possible when classification algorithms take into account qualities with a large number of possible values.
Mohan et.al, 2019	Random Forest (RF) and Linear Method (LM)	It Can be used for regression and classification problem and solve over fitting a problem in the decision tree	Need much time for training
Khourdifi et al, 2019	Fast Correlation-Based Feature Selection (FCBF)	Reducing over-fitting. Making decisions based on noise is reduced when there is less redundant data.	When the number of variables is enormous, the amount of time it takes to compute is very significant.
Reddy et al, 2019	Random Forest	Easy to configure and use	higher likelihood of overfitting, model harder to explain and costlier to maintain

Author & year	Method	Merits	Demerits
Tougui et al, 2020	Artificial Neural Network	This study proposes Matlab as best ML tool and ANN as best classifier among six classifiers	Not an exhaustive study of tools and classifiers as only six tools and techniques were compared
Badriyah et.al, 2020	Random Forest	when compared to other categorization algorithms, has the best accuracy	No tests of the precision of classification procedures employing the optimization's default parameter value have been performed.
Li et al, 2020	ANN and Fuzzy Logic	Its use in healthcare settings for diagnosing cardiac disease is straightforward.	More execution time is required to generate the result.
Singh et al, 2020	Position-Specific Mutation (PSM) and One-Hot Encoding (OHE) technique	The PSM approach may also be useful for forecasting the fatality or severity of diseases caused by other mutations.	Predictions of severity could use more effort to improve their accuracy.
Kilicarslan et al, 2021	GA-CNN (Genetic Algorithm and Convolutional Neural	Analysis of medical data by these models aids in diagnosis and	High complexity

Author & year	Method	Merits	Demerits
	Networks) and GA-SAE (Genetic Algorithm and Stacked Autoencoder)	reduces the likelihood of mistakes brought on by human inexperience or exhaustion.	
Lee et al, 2021	synthetic minority over-sampling technique	It's not making copies, but rather creating simulated data points that are very similar to the real ones but with some little differences.	It is imbalanced and has limited features
Gopal et al, 2021	correlation coefficient function	Whether the classes are uniformly distributed or not, a high quality score requires accurate classification of a large fraction of negative data instances and a significant fraction of positive data instances.	It has a limitation of specific number of dataset.
Archika et al, 2022	Correlation Based Feature Selection (CFS) and Gain Ratio	High-correlation features usually always have the same impact on the dependent variable because of their linear dependence.	It takes into account attributes with several possible values, which could cause overfitting.

2.2.3 Study on Optimization of classifiers for disease classification

Heart disease diagnosis with an optimization algorithm can be fruitful in terms of higher accuracy and sensitivity. Finding an acceptable optimal solution among multiple solutions for a specific problem is known as optimization. The following works have used various optimization techniques to improve the performance of their models.

Gadekallu et al. (2017) created a system for the diagnosis of cardiovascular and metabolic diseases using a cuckoo search optimized rough sets based attribute reduction and fuzzy logic system. First, cuckoo search and rough set theory are used to reduce the number of features that need to be examined in order to make a disease prediction. As a result of the first procedure, the computational load of the fuzzy logic system is decreased, and its performance is improved. In the next phase, disease datasets are categorized using fuzzy rules and membership functions. The experimental findings validate the superiority of the suggested algorithm above the current best practices. Its convergence rate is poor, and it tends to become stuck on local maxima.

Heart disease was detected using a variety of classification algorithms in a study by Wijaya et al. (2018), including NB, DT, and k-NN. In this experiment, PSO is used as a feature selection strategy to boost accuracy. Therefore, the PSO feature selection approach can be used to enhance the classifiers' performance accuracy. It is intended that the classification results will aid in the accurate diagnosis of heart condition by doctors and other medical experts, resulting in a decrease in the mortality rate caused by cardiovascular disease. However, it has a low convergence rate in the iterative process andis prone to reaching a local optimum in high-dimensional spaces.

Prabhakar et al. (2020) proposed a method to classify schizophrenia based on EEG data utilizing feature extraction, optimization methods, and suitable classifiers. Three feature extraction techniques are used, including Partial Least Squares (PLS) Nonlinear Regression, Expectation Maximization based Principal Component Analysis (EM-PCA), and Isometric Mapping (Isomap). Four optimization techniques are including in the Flower Pollination algorithm, the Eagle strategy with a variety of evolution algorithms, the obtained characteristics are then refined using the Backtracking search optimization technique and the Group search optimization algorithm. Adaboost (AdaB)and NB Classifier variants are used to categorize the optimal values.

A BC prediction system (BCPS) based on an OANN was proposed by Supriya et al. in 2020. The raw BC information is the primary input. The data is preprocessed using normalization and RMA to fill in any blanks. After that, the features are often selected with the help of the MDF algorithm (Modified Dragonfly). The features are then fed into a categorization system. In this case, OANN is used to categorize the features. The best

outcome is reached by employing the Gray Wolf Optimization (GWO) method. GWO's slow convergence rate, inaccurate solutions, and ease of getting stuck in a local optimum all make it a poor choice.

Parameter Free BAT (PF-BAT) optimized Fuzzy K-nearest neighbor (PF-FKNN) classifier is the basis of the expert model proposed by Kaur et al. (2021) for the detection of new coronaviruses. The proposed model retrieves feature from a MobileNetv2 that was learned via transfer learning, and then trains using FKNN. Experimental results on gold-standard COVID CT scan data demonstrate that the proposed algorithm improves upon prior state-of-the-art approaches in terms of validation accuracy. The biggest problem is Large datasets can slow down the prediction phase.

Wazery et al. (2021) developed an efficient classification technique that combines Opposition-Based learning (OBL) with the slime mould algorithm (SMA) based on kNN, with the goal of reducing the amount of time spent on feature selection (FS) and classification, they dubbed this method ISMA-kNN. ISMA is an algorithm designed to address the shortcomings of the original SMA, such as its tendency to become stuck in local minimum regions and its imbalance between exploitation and exploration, especiallywhen dealing with issues of high dimension. ISMA is a randomized optimization method, hence its limitations could be different each time it is run. Users may become confused if they expect to see the same subset of features they chose in one run appear in another.

The proposed model by Nawaz et al. (2021) improves the accuracy, precision, sensitivity, and specificity with which cardiac disease can be diagnosed. It is expected that the suggested Expert System will be simple to operate. The prevalence of cardiac disease has been brought to light by recent studies. The current difficulties experienced by medical professionals and patients while diagnosing cardiac illness are thought to be better addressed by the creation of a system based on Gradient Descent Optimization (GDO). However, there are cases where a model converges to a state that isn't optimal dueto a steady error gradient.

To enhance the precision of the baseline CNN model, Nainwal et al. (2022) developed a method as an optimization methodology. In this case, we optimize using the refined version of the Monarch Butterfly Optimization (MBO) method. ECG data were analyzed for morphological aspects and features based on wavelet coefficients. DR is

achieved by the use of an IMBO (Improved Monarch Butterfly optimization) algorithm on the feature vector. Convolution neural networks use these refined characteristics for signal classification. Overall the migrating populations are less than half the size they need to be to avoid extinction.

Cenitta et al. (2022) proposed Ischemic Heart Disease Squirrel Search Optimization (IHDSSO) model in conjunction with Random Forest classifier, which ensures better feature selection and accuracy with respect to other state-of-the-art optimization algorithms.

Table 2. 3 Comparison of Optimization of classifiers for disease classification

Author & year	Method	Merits	Demerits
Gadekallu et al, 2017	Cuckoo search optimized rough sets based attribute reduction and fuzzy logic system.	Simpler implementation and fewer adjusting factors.	It has a low convergence rate and is prone to becoming stuck in locally optimal solutions.
Dolatabadi et al, 2017	Support Vector Machine (SVM)	It is possible to employ mobile systems to transmit heart signals and monitor patients remotely online.	Its computational time is high.
Wijaya et al, 2018	Particle Swarm Optimization	It reduces the mortality rate caused by heart diseases.	Due to the high dimensionality of the space and the slow convergence rate of the

Author & year	Method	Merits	Demerits
			iterative technique, it is easy to get stuck in a local optimum.
Prabhakar et al, 2020	• Partial Least Squares (PLS) Nonlinear Regression technique, • Expectation Maximization based Principal Component Analysis (EM-PCA) technique • Isometric Mapping (Isomap) technique.	Nonlinear degrees of freedom underlying complicated natural data can be discovered.	Technical implementations often assume no missing values
Supriya et al, 2020	Gray Wolf Optimization (GWO) algorithm	fewer moving parts, less complicated principles, and straightforward application	The convergence rateof GWO is sluggish, the solution accuracy is low, and it is simple to get stuck in a local optimum.
Kaur et al, 2021	Parameter Free BAT (PF-BAT) optimized Fuzzy K-nearest neighbor (PF-FKNN) classifier	It can be used to solve problems involving regression and classification.	With large data, the prediction stage might be slow.
Wazery et al, 2021	ISMA	ISMA is computationally simpler than the	Since ISMA relies on randomization for its optimization, the

Author & year	Method	Merits	Demerits
		original SMA and can produce more effective solutions.	results may vary depending on the iteration.
Nawaz et al, 2021	Gradient Descent Optimization	Convergence is more stable than Stochastic Gradient Descent	If the error gradient is constant, the model may converge to a state that is suboptimal.
Nainwal et al, 2022	Improved Monarch Butterfly optimization (IMBO) algorithm	Cut down on production costs drastically while keeping efficiency levels high.	Overall the migrating populations are lessthan half the size they need to be to avoid extinction.
Cenitta et al, 2022	Ischemic Heart Disease Squirrel Search Optimization (IHDSSO), with Random Forest classifier	High performance compared to existing models	Improvements required in terms of convergence accuracy and convergence speed.

2.2.4 Study on Ensemble Classification for Diseases prediction

By combining dataset deconstruction with ensemble learning, Prince et al.(2018) provide a novel approach to dealing with missing data at the source. The study found that the DNN's accuracy is comparable to that of conventional feature-based classifiers. This suggests that the feature set, rather than the classification method, is the limiting element during the classification process. The provided work assumes that all tests were passed on the first try, which is not always the case. While this is largely the case, several of the experiments performed here were completely meaningless noise.

Using a medical dataset, Latha et al. (2019) demonstrate the method's usefulness in early disease prediction by focusing on two main areas: First, the method needs to be implemented, and second, the accuracy of weak classification systems needs to be

improved. The ensemble algorithms bagging, boosting, stacking, and majority voting were all used in the experiments. The Brute-force method is used to choose features in this research. The feature selection techniques helped the ensemble algorithms perform better.

In order to better predict who may develop Alzheimer's disease in the future, Rohini et al. (2019) propose conducting a study to develop a learning algorithm. A person's chance of getting Alzheimer's disease can also be assessed by looking at the characteristics of those diagnosed with Mild Cognitive impairment (MCI) or Pre-Mild Cognitive Impairment (Pre-MCI). In the current research, three different classifier models—the KNN classifier, the SVM classifier, and the Gaussian NB classifier model— are applied, and a novel classifier is proposed by integrating the classifiers in a single model. All classifier models can benefit from using more optimized processes than the currently implemented gradient descent, such as the conjugate gradient, BFGS, and L- BFGS.

Abdar et al. (2019) built their NE-nu-SVC model with linear, polynomial, RBF, and sigmoid kernels. Combining nu-SVC with other powerful ML techniques within a NE model, like stochastic gradient descent (SGD), sequential minimum optimization (SMO), recursive function (RF), neural network (NB), staking, bagging, and voting, was done. In addition, the effectiveness of the new model was enhanced by the application of data balancing and features selection. The Cleveland and Z-Alizadeh Sani data sets are used. This approach has the drawback of being quite computationally intensive.

Because the early phases of DR are so poorly characterized by current models, Qummar et al. (2019) set out to remedy this problem. To detect and classify the various stages of DR in color fundus images, they proposed a CNN ensemble-based system. The ensemble method is a class of meta-algorithms that integrates different types of ML into a unified predictive framework. The variance (Bagging), bias (boosting), and prediction quality (stacking) can all be reduced with its help. Stacking is a method for creating a new prediction model by combining the results of many existing models. The results show that the proposed ensemble model not only identifies all DR phases, but also outperforms earlier state-of-the-art methods. However, it is vulnerable to anomalies because each successive classifier is responsible for correcting the mistakes made by its predecessors.

Disease prediction model (DPM) for T2DM and hypertension developed by Fitriyani et al. (2019) using isolation forest (iForest), SMOTETomek, and ensemble learning. The iForest was used to find and eliminate outliers in the dataset, the SMOTETomek was employed to normalize the data, and ensemble learning was utilized for disease forecasting. The proposed DPM was also incorporated into a smartphone app that makes use of IoT-based sensors to provide a diagnostic in real time. As a result, the DPM was put into action to check for hypertension and/or type 2 diabetes, with the results being sent to the user's mobile-app. However, this approach fails to scale well for data with several dimensions.

To combine data from many sources and leverage the "wisdom of experts," An et al. (2020) proposes a deep ensemble learning framework that makes use of deep learning algorithms. Two sparse autoencoders are trained for feature learning in the voting layer to reduce attribute correlation and ultimately diversify the base classifiers. The neural network is useful as a Meta classifier. By calculating similarities between an ensemble of probabilistic forecasts, we can generate improved forecasts. The proposed deep ensemble learning approach has potential applications, including Alzheimer's disease classification. It may be difficult or expensive to get labels that are both objective and accurate.

In order to categorize the ECG signal, Sun et al. (2020) recommended using a multi-label classifier. There are bounds that each type of multi-label classifier must obey. The use of an ensemble multi-label classifier is recommended as a means to improve classification performance for ECG signal classification. The approach relies on many multi-label classifiers as its primary classifier. Each of the primary classifiers is given a weight based on their mutual information. The ensemble classification results are based oncomparisons to the threshold values for each class. Using the F1 score as the objective function, the best threshold values are determined using the GA technique. The WFDB toolbox is first used to extract standard electrocardiogram (ECG) parameters such as heart rate, signal quality, HRV, QRS, morphology, P-wave, QT, and so on. They can be difficult to debug and can be computationally expensive.

Brain tumors and neoplasms, including glioma, meningioma, and pituitary adenoma, as well as auto-immune disease lesions like MS, can be differentiated using magnetic resonance imaging (MRI), as described by Shafi et al. in 2021. The learner is

built on a support vector machine (SVM) classifier and a majority voting prediction model. Image preprocessing is required on the quantized volumes before statistical texture-features such the gray-level contrast matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size zone matrix (GLSZM), and neighborhood gray-tone difference matrix (NGTDM) can be extracted. The proposed model is more imperceptible and resilient than previous approaches to similar problems in image processing. The results indicate that the host image suffers just a negligible loss of quality.

Dhar et al. (2021) proposed a unique ensemble model to aid clinicians in the early detection of Chronic Obstructive Pulmonary Disease (COPD) and the timely administration of suitable medicines. The MSEN model, an improved version of the weighted voting-based Multistage Ensemble model. This research uses a weighted voting method to integrate the generated ensemble models into a single MSEN model. Each classifier in every pool has its hyperparameters optimized using the GA. The weights of two newly constructed ensemble models and each classifier are optimized using a grid search. But, it has high time consuming.

In order to adopt the best classification outcomes for cervical cancer prediction, Ilyas et al. (2021) used an ensemble classification approach based on a substantial voting procedure. The analysis employed a wide variety of classifiers, including DT, SVM, RF, KNN, NB, MP, J48 Trees, and LR. But, it can be more difficult to interpret.

Essa et al. (2021) proposed a multi-model, deep learning-based strategy for ECG data classification. They present two deep learning bagging models for identifying and categorizing arrhythmias from recorded heartbeats. The first model (CNN-LSTM) is based on a combination of a convolutional neural network (CNN) and a long short-term memory (LSTM) network, which allows it to capture both the local characteristics and the temporal dynamics in the ECG data. The second model (RRHOS-LSTM) combines some classical traits, like RR intervals and higher-order statistics (HOS), with the LSTM model to effectively identify clusters of aberrant heartbeats. However, they are more difficult to implement and need more data for training than regular RNNs.

The CNN method was proposed by Gamal et al. in 2022 as a novel strategy for categorizing illness phases. To begin, they sorted through the MRI scans and divided them into a secure manner. After that, a quick and easy preprocessing pipeline was applied

without the need for registration to the dataset. There were several experiments done to compare the efficacy of various 3D categorization structures. At last, the best models are put through an ensemble learning process. Using an extended version of the ADNI dataset, the superior performance of the suggested technique was proved. But, it required to test large data set.

De Jesus et al. (2023) introduced EnsembleDVX, which uses a combination of three separate CNNs models to aid in the diagnosis of COVID-19 using CT scans. Thebest three CNNs were initially selected using the k-fold resampling method. According to the findings, DenseNet169, VGG16, and Xception were the most effective models. The next step was to figure out how to pool the models' individual classifications. As seen in the findings, ensemble models are more likely to produce accurate predictions. However,a lot of training data is necessary for the CNN to be effective.

Table 2.4 Comparison of Ensemble Classification Methods for Diseases Detection

Author & year	Method	Merits	Demerits
Prince et al, 2018	Deep Neural Network (DNN)	Its ability to execute feature engineering by itself.	Some of the tests performed for this study amount to nothing except background noise.
Latha et al, 2019	Brute force method	It iterates over the array many times in order to get every possible solution	Exceedingly slow as they may have to perform every possible combination of characters before they achieve their target.
Rohini et al, 2019	KNN, SVM, and Gaussian NB classifier model	The performance metrics was good	Expensive cost

Author & year	Method	Merits	Demerits
Abdar et al, 2019	Genetic Algorithm	It provides multiple optimal solution	They can be computationally expensive
Qummar et al, 2019	Bagging, boosting	It can be used with several hyper-parameter tuning options to improve fitting	The requirement that each successive classifier correct the mistakes made by the one before it makes the system vulnerable to outliers.
Fitriyani et al, 2019	synthetic minority oversampling technique tomek link (SMOTETomek)	Rather than making copies, it's producing synthetic data points that differ somewhat from the real ones.	When working with large datasets, SMOTE's limitations become apparent.
An et al, 2020	Deep Ensemble Learning Framework	The performance matrix is good.	It may be difficult or expensive to get labels that are both objective and accurate.
Sun et al, 2020	genetic algorithm method	They usually perform better than traditional feature selection techniques	They can be difficult to debug and can be computationally expensive.
Shafi et al, 2021	SVM	It decreases both prediction mistake and generalization blunders.	The results indicate that the host image suffers just a negligible loss of quality.

Author & year	Method	Merits	Demerits
Dhar et al, 2021	Grid search technique	The ideal values for the hyperparameters will be calculated based on the training set data.	Prolonged effort required
Ilyas et al, 2021	ensemble classification method	Better performance and more accurate predictions are possible with an ensemble than with any one model working alone.	It can be more difficult to interpret.
Essa et al, 2021	Convolutional neural network (CNN) and long short-term memory (LSTM) network.	It is capable of timing analysis and abstract feature extraction simultaneously.	They are more involved than regular RNNs and need more information to learn.
Gamal et al, 2022	Convolutional Neural Network (CNN)	The network's classification and accuracy as a whole can be enhanced by selecting appropriate parameter values.	It has a required to test large data set
De Jesus et al, 2023	Convolutional neural network (CNN)	do not need human oversight to accomplish the task of finding crucial characteristics	In order for the CNN to be useful, significant amounts of training data are required.

2.3 RESEARCH GAP

Many ML models are developed and validated on a specific dataset without much validation on different datasets, leading to concerns about the model's generalizability. Only few studies validate their models in real-world clinical settings, which is crucial for assessing their practical utility. There is a need for more research on interpretable models or methods to explain black-box models in the context of heart disease prediction.

This provides scope to build robust heart disease prediction models with improved performance and generalizability, while it is essential to understand the type of features in the dataset and perform proper pre-processing, balance the dataset if it is imbalanced, recognize the relevant features in the dataset which highly contribute to the disease prediction, integrate the best diverse classifiers to form strong ensemble models with generalizability and less complexity. Detecting heart disease is a challenging task because misdiagnosis can result in serious problems.

2.4 CHAPTER SUMMARY

In this chapter, an in-depth survey of existing models used to classify human diseases and specifically heart disease is done by segregating the literature survey into Review of Feature Selection Methods, review of Machine Learning techniques, Study of Optimization methods and review of ensemble classifiers. Machine Learning Techniques that focus on normalizing the data, balance the data, perform feature selection which would save time significantly and improve the prediction results and model generalizability are always sought after.

Research on prediction of heart diseases risk can be done by building models firstly, by incorporating techniques to identify the most significant features on heart dataset which enhances the classifiers' performance and secondly, devising robust interpretable ensemble classifiers to integrate the strengths of single classifiers, and balance the bias and variance.