TAYLOR-BIRD SWARM OPTIMIZATION-BASED DEEP BELIEF NETWORK FOR MEDICAL DATA CLASSIFICATION

ALHASSAN AFNAN MOHAMMED

UNIVERSITI SAINS MALAYSIA

2022

TAYLOR-BIRD SWARM OPTIMIZATION-BASED DEEP BELIEF NETWORK FOR MEDICAL DATA CLASSIFICATION

by

ALHASSAN AFNAN MOHAMMED

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

August 2022

ACKNOWLEDGEMENT

In the name of Allah, Most Gracious, Most Merciful. Praise be to Allah for His mercy which has enabled me possible to complete this thesis.

I would like to express my heartfelt gratitude to my supervisor Dr. Wan Mohd Nazmee Wan Zainon who guided me throughout this thesis and provided me with his irreplaceable support.

I will always be grateful for my parents, husband, daughter, brothers, sisters, and friends I wouldn't have been able to get through everything without their support.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ALGORITHMS

Page

LIST OF ABBREVIATIONS

BSA	Bird Swarm Algorithm
DBN	Deep Belief Network
FCM	Fuzzy C-Means
SaaS	Software-as-a-Service
XaaS	Everything-as-a-Service
IaaS	Infrastructure-as-a-Service
PaaS	Platform-as-a-Service
ABE	Attribute-based encryption
CDSS	Clinical Decision Support Systems
CVD	Cardiovascular disease
SVM	Support vector machine
KNN	K-nearest neighbour
EP	Evolutionary Programming
ANN	Artificial neural network
PNN	Probabilistic neural network
MLP	Multilayer perceptron network
LVQ	Learning vector quantization
DTH	Digital twin healthcare
CIW	Class influence weight
PPDP	Privacy-preserving disease prediction
RBBB	Right Bundle Branch Block
RCNN	Recurrent convolutional neural network
PDPSS	Privacy-Aware Disease Prediction Support System
CART	Classification and regression tree

LPP	Locality Preserving Projections
CBR	Case-Base Reasoning
PROCAM	Prospective Cardiovascular Munster
GCS	Google cloud storage
EFS	Ensemble Feature Selection
FS	Feature Selection
SI	Swarm Intelligence

RANGKAIAN KEPERCAYAAN MENDALAM BERASASKAN PENGOPTIMUMAN KAWANAN TAYLOR-BIRD UNTUK KLASIFIKASI DATA PERUBATAN

ABSTRAK

Data perubatan menunjukkan ciri-ciri tertentu untuk menjadikan klasifikasi lebih baik dalam pelbagai contoh penyelidikan. Kini terdapat banyak pendekatan klasifikasi data perubatan, dari mana prognosis dan diagnosis perubatan biasa diperolehi. Klasifikasi data perubatan adalah rumit tetapi mencari strategi yang paling tepat untuk masalah klasifikasi perubatan bersama dengan parameter optimumnya tidaklah terlalu sukar. Sumbangan utama penyelidikan ini adalah klasifikasi data perubatan menggunakan rangkaian kepercayaan yang mendalam berasaskan pengoptimuman kawanan Taylor-Bird (DBN berasaskan Taylor-BSA). Pada mulanya, pra-pemprosesan data perubatan dilakukan dengan menggunakan transformasi log yang menukar data ke julat nilai seragamnya. Kemudian, proses pemilihan ciri dilakukan dengan menggunakan fuzzy-c-means (FCM) yang jarang untuk memilih ciri penting untuk mengklasifikasikan data perubatan. Menggabungkan FCM jarang untuk proses pemilihan fitur memberikan lebih banyak manfaat untuk menafsirkan model, kerana teknik jarang ini memberikan fitur penting untuk pengesanan, dan dapat digunakan untuk menangani data dimensi tinggi. Kemudian, ciri-ciri yang dipilih diberikan kepada jaringan kepercayaan mendalam (DBN), yang dilatih menggunakan algoritma kawanan burung berdasarkan Taylor (Taylor-BSA) yang dicadangkan untuk pengesanan. Di sini, Taylor-BSA yang dicadangkan direka dengan menggabungkan siri Taylor dan algoritma kawanan burung (BSA). Taylor-BSA – DBN yang

dicadangkan mengungguli kaedah lain, dengan ketepatan maksimum 93.4%, kepekaan maksimum 95%, dan kekhususan maksimum masing-masing 90.3%.

TAYLOR-BIRD SWARM OPTIMIZATION-BASED DEEP BELIEF NETWORK FOR MEDICAL DATA CLASSIFICATION

ABSTRACT

Heart disease classification is considered a challenging and complex task in the field of medical informatics. Various medical data classification methods are developed in the existing research works, but achieving higher classification accuracy is a great challenge in the medical sector due to the presence of noisy, and high-dimensional data. Fuzzy clustering-based filtering methods are introduced for essential feature selection. From the selected features, deep learning has become an important stage for disease diagnosis. However, finding the most appropriate deep learning algorithm for a medical classification problem along with its optimal parameters becomes a difficult task. Deep Belief Network (DBN) is a sophisticated learning system that requires a high level of approach and executes well. The major contribution of this research is to introduce a Taylor-Bird Swarm optimization-based Deep Belief Network (Taylor-BSA-based DBN) for medical data classification. Firstly, the pre-processing of medical data is done using log-transformation that converts the data to its uniform value range. Then, the feature selection process is performed using sparse fuzzy-c-means (FCM) for selecting significant features to classify medical data. Incorporating sparse FCM for the feature selection process provides more benefits for interpreting the models, as this sparse technique provides important features for detection and can be utilized for handling high-dimensional data. Then, the selected features are given as input to the DBN classifier which is trained using the Taylor-based bird swarm algorithm (Taylor-BSA). Taylor-BSA is designed by combining the Taylor series and bird swarm algorithm (BSA). The proposed Taylor-BSA–DBN

outperformed other methods, with the highest results accuracy of 93.4%, the sensitivity of 95%, and specificity of 90.3%, respectively.

CHAPTER 1

INTRODUCTION

1.1 Introduction

According to the World Health Organization (WHO) report, heart disease is the main reason of death globally with 18 million people dying every year (WHO, 2020). Clinical judgments are frequently based primarily on clinicians' intuition and experience instead of the dataset's information-rich data. This approach causes unintended biases, mistakes, and even exorbitant medical expenses, all of which have an impact on the quality of care given to affected individuals (Lashari, Ibrahim, Senan, &Taujuddin, 2018).

The data mining approach holds promise since data modelling and analysis technologies can create an information-rich atmosphere that can dramatically increase the performance of healthcare judgments. Effective data mining techniques have incentivised all parties participating in medical-related companies to fully employ them since they have understood that data mining is critical in gathering critical data for the entire sectors concerned. Disease recognition is now one of the uses of data mining. To accomplish this, one requires a clinical database to uncover hidden patterns and then extort relevant information from the clinical dataset (Sharma, Singh, & Khatri, 2016).

1.2 Medical Data Mining

The detection, and treatment, including preventing disease, injury, as well as other physical and mental disabilities in individuals, are all covered by healthcare. In many places, the medical business is quickly changing. Because they produce huge

volumes of data, such as electronic health records, administrative records, as well as other standard findings, the healthcare company can be defined as a data-rich environment (Abdeldjouad, Brahami, & MattaEmail, 2020).

Nevertheless, the medical data are underutilised (Sen, Patel, & Shukla, 2013). Data mining can sift through these massive amounts of information in search of new and useful data. In the medical field, data mining is mostly employed to anticipate different illnesses and help clinicians make clinical decisions. Whether data mining is utilised in the medical industry, the main goal is to uncover meaningful and intelligible patterns by analysing enormous volumes of information (Sharma, Singh, & Khatri, 2016). The data patterns aid in the prediction of business or data trends, as well as determining what to do in response. Data mining could be employed in the medical industry to lower expenses by enhancing effectiveness, enhancing individual's life quality, recognise treatment plans as well as best practices, measuring efficiency, and clinical claims, and, maybe most likely, save the livelihoods of many individuals, all of which enhances the level of patient care.

This data has a reputation for being complicated to understand. Moreover, data mining ideas can analyse and classify massive amounts of data, group variables with the same behaviours, and forecast upcoming events, among other benefits for monitoring and managing medical systems that are constantly attempting to protect individuals' confidentiality (Kolling, Furstenau, Sott, Rabaioli, Ulmi,& Bragazzi, Tedesco, 2021). In research and healthcare, predictive data mining is now a significant analytical tool.

1.3 Deep Learning Based Classification

Machine learning has a branch called deep learning (Grossfeld, 2017). Deep learning can be used to do classification. Deep learning is a subset of machine learning that builds an "Artificial Neural Network (ANN)", which can train and make smart judgments on its own by layering techniques. Since both belong underneath the umbrella of artificial intelligence, deep learning is the engine that drives the most human-like AI. (Reference Figure 1.1 for further information.)

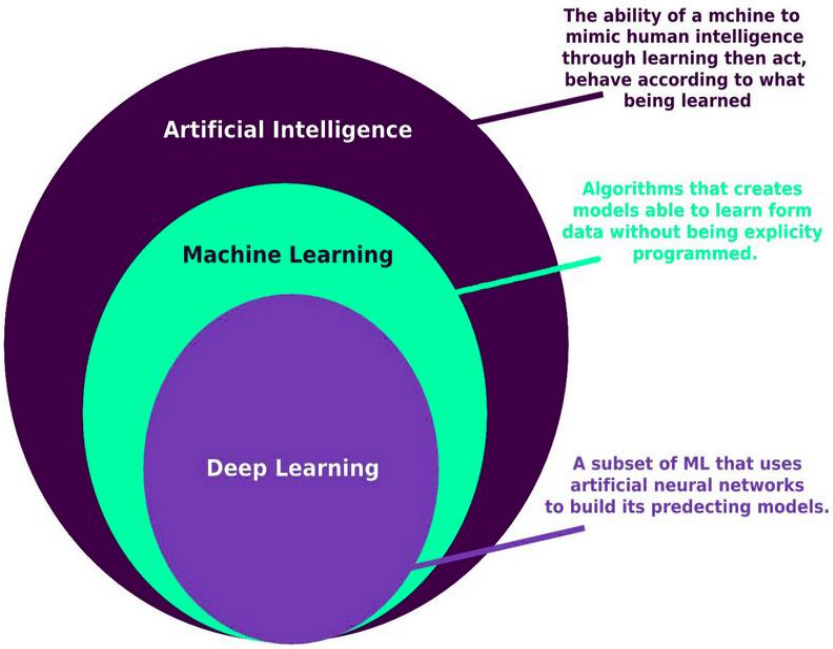

Figure 1.1 Deep Learning Overview (Source: Imam, 2020)

In the area of machine learning, deep learning is a relatively recent field. Its algorithms generate multilayer designs with hierarchical interpretations of the incoming data on the go. Abstract notions are represented through the model's higher-level forms, specified as a non-linear combination of lower-level interpretations. The numerous contextual aspects essential for categorisation are amplified by these different attributes (Pinaya, Gadelha, Doyle, Noto, Zugman, Cordeiro, Jackowski, Bressan, & Sato, 2016).

1.3.1 Classification Techniques for Decision Making

Most of the techniques are implemented using practical tools for accelerating the process to make decision-making effective. Hospitals and healthcare centres directly affect the health and well-being of the community. Whenever the disease diagnosis is considered, several techniques are devised, which lead to effective decision-making and related activities. Meanwhile, the available information is vast, and hence, accurate decision-making becomes complex (Liu, Lu, Ma, Chen, & Qin, 2016).

In clinical coding, a medical classification is employed to convert descriptions of clinical treatments or treatments into standardised statistical code. Diagnosis classification is a set of clinical features employed to monitor diseases and other health issues, like diabetes mellitus and heart disease, as well as infectious diseases including norovirus, the flu, and athlete’s foot. Health care practitioners, government health programmes, private insurance providers, workers' compensation carriers, software developers, and many others utilise these diagnosis and process codes for several applications in healthcare, public health, and clinical informatics.

Clinical information has previously been effectively employed to construct several Clinical Decision Support Systems (CDSSs), all of which have a favourable effect on practitioner efficiency and medical procedure (Sutton et al., 2020). Moreover, CDSS can improve the accuracy of diagnosis by reducing the time taken for diagnosis. Thus, classification can be used to extract essential information for improving CDSS using a large amount of clinical data (Liu et al., 2016). Additionally, categorisation is an integral part of data mining tasks. The classification aims to devise a classification

model for predicting diseases to enable effective decision-making for physicians regarding diagnoses.

Clinical Decision Support Systems (CDSS) are a kind of software devised for clinical treatments and performing diagnoses in a cloud computing infrastructure. This system applies the clinical knowledge for diagnosing the diseases and describes the medical recommendations for the patients. These systems are developed for practitioners, and the conditions are diagnosed using empirical experiences for helping medical experts.

Clinical decision support (CDS) integrates the knowledge of the medical domain concerning the healthcare data to improve decisions regarding patient healthcare. The usage of CDS systems ranges from reminders for providing safety alerts to public health notifications. CDS are responsible for providing effective decisions in health care systems. In addition, CDSS can support the improvements in terms of quality, efficiency, and safety to become a global priority in health care systems. CDSS utilise a sophisticated data mining mechanism for helping clinicians to make effective decisions and had gained significant attention (Nazari, Fallah, Kazemipoor, & Salehipour, 2018).

1.4 Motivation

Disease prediction predicts the disease by extracting the information from the medical data, which is of great interest in the growing research. Many researchers tried their research work on this topic, and thus, there exist several contributions to this topic. This research work is inspired by the recent advancements in this machine learning techniques and clustering techniques. One can predict the disease at different levels as normal, abnormal, level 1, level 2 or serious. In several reviews, people

express diverse beliefs about disease prediction. Thus, feature-based disease prediction would be the most suitable choice for this research. Some of the classical techniques have presented an overview of disease prediction systems.

Further, the current version of disease prediction methods is mostly categorised into three types, content-based, collaborative and hybrid disease prediction methods. However, there exist many limitations to these methods in terms of accuracy. Possible extensions are needed that can improve disease prediction capabilities as well as make disease prediction systems applicable to a broader range of applications.

Medical data has been widely used for disease prediction, and there are so many datasets to support the research in medical data classification. Some authors have used the datasets and proposed a disease prediction model based on Naive Bayes, Support Vector Machine and deep learning methods. By using standard machine learning techniques, the contributed works have used the naive Bayes and SVM algorithms, which are incorporated into the domain of disease prediction systems to automatically classify the disease as normal or abnormal. The effects of classification performance are discussed by many researchers. However, many techniques for predicting the disease had resulted in poor accuracy. Taking accuracy as a major concern, this research utilised a log transformation procedure for reducing the complexity thereby, increasing the classification accuracy.

1.5 Problem Statement

The medical area produces increasingly voluminous amounts of data which are becoming more complicated. The produced medical data have certain characteristics that make their analysis very challenging and attractive. Medical data costs a lot of funds and it is extremely costly to run trials and to find patients agreeable to help,

especially for rare diseases. In recent years, the development of computer technology has made great progress in data mining and machine learning technology. Supporting massive amounts of data and using machine learning algorithms to utilize the data effectively can enhance the value of data. At present, heart disease, diabetes, and cervical cancer is the cause of death worldwide. Machine learning methods for the classification of the medical dataset in the existing research works don't provide higher results due to the presence of noisy, and high-dimensional data.

In the high-dimensional setting, data collection is often prone to noise. Often the high dimensionality might still prevent proper estimation of the statistics of the noise itself. For high-dimensional datasets, there is the so-called curse of dimensionality. The searchable volume in the hyperspace becomes small, compared with the vast feasible search space. Thus, any solution procedure can only sample a subset of sparse points with essentially zero sampling volume to make sense of the vast datasets. Thus, it is a huge challenge with an almost impossible task for finding global optimality.

As the number of dimensions increases, the number of features also increases, often far more rapidly, which means that there is huge sparsity associated with such high-dimensional features. In addition, some correlation may exist between different dimensions, and thus features can be difficult to define. Therefore, the accuracy of prediction still needs to be improved (Bashir et al., 2019). Thus, having a feature selection is compulsory to gain the best possible accuracy in predicting heart disease. Fuzzy clustering-based filtering methods are introduced for essential feature selection (Liu et al., 2012). From the selected features, the classifier has become an important stage for disease diagnosis.

Deep Belief Network (DBN) is a sophisticated artificial neural network-based learning system that requires a high level of approach and executes well (Kim et al., 2017). Also, with the current emergence of structured data, DBN with an increased number of application-specific hidden units has shown tremendous improvement. Though, in the process of parameter selection for the DBN classifier, network parameters are still modified according to experience. At this time, the diagnosis model has the difficulty of inadequate constancy and high randomness of diagnosis. Thus optimisation methods have been introduced to solve this issue which improves the system's performance (Shukla, 2020; Ali et al., 2020). The proposed work will focus on a clustering-based feature selection algorithm and a Taylor-based bird swarm algorithm (Taylor-BSA) is introduced for heart disease detection.

1.6 Research Questions

The section demonstrates various research questions related to the methods of medical data categorisation.

- Whether the developed training algorithm for the Deep Belief Networks performed effective disease prediction and whether the devised hybrid classifier can offer improved medical data classification accuracy for disease prediction over any other existing schemes?

- Does the modelled medical data classification system achieve high classification accuracy and enhanced performance?

- How will the accurate categorisation of clinical data be enabled in the developed clinical data classification system, and will the developed method meet the robust performance?

1.7 Research Objectives

This research aims at analysing the medical data for disease prediction using several approaches. The objectives of the research can be explained as follows:

- ❖ To predict heart disease and its severity using a classifier based on a deep belief network.
- ❖ To optimize this classifier using the bird swarm algorithm.
- ❖ To test the proposed methodology and compare the results with existing algorithms on the heart dataset.

1.8 Contributions of the Thesis

The major contribution of the work is to design a Taylor-BSA-based DBN classifier to acquire accurate classification results. The proposed Taylor-BSA is obtained by modifying the update equation of the BSA algorithm with the prediction method, the Taylor series. Thus, integrating Taylor with BSA enhances the convergence rate and provides an optimal solution for performing the medical data classification for predicting the disease. Moreover, Taylor exhibits a lot of merits, like formulating a solution based on the previous records. Thus, interpreting the Taylor series in BSA interprets the previous best solutions for finalising the best solution, which are the optimal weights for tuning the DBN classifier.

1.9 Proposed Methodology

The primary intention of this research is disease prediction using the medical records of the patients. The proposed method is processed using three phases, which involve Preprocessing, Feature Selection, and Classification. Initially, clinical data is

subjected to the preprocessing phase for making highly skewed distributions and less skewed distributions using log transformation. Thus, log transformation is employed for generating the patterns using more interpretable and helps fulfil the supposition, thereby reducing the skew and normalising the data. Then, for the feature selection procedure, sparse Fuzzy c-means (FCM) is used to examine the important features for categorising the health information.

A smart feature selection process should be able to reduce computing complexity while increasing classification performance. As a result, using sparse FCM for the feature selection yields additional advantages for analysing the systems, as this sparse system provides key classification features and may be used to handle high-dimensional data. As a result, the chosen characteristics were fed into the Deep Belief Network (DBN), which was trained for categorisation by using the suggested Taylor-based bird swarm method (Taylor-BSA). The suggested Taylor-BSA is created by merging the Taylor series with the Bird swarm method in this study.

1.10 Scope

In recent years, diagnosis of the disease has become a notable field, since it focuses on the early prevention of the diseases such as the heart. Heart disease explains a variety of conditions with the purpose of influencing your heart. Today, these diseases are the leading cause of death global with 17.9 million deaths annually, as per the World Health Organization reports by 2020. A variety of unhealthy activities are the cause of the raise in the risk of heart disease similar to high cholesterol, obesity, increase in triglycerides levels, hypertension, etc... There are definite signs which the American Heart Association lists similar to the persons having sleep issues, a definite raise and reduction in heart rate (irregular heartbeat), swollen legs, and in some cases

weight gain happening moderately fast; it can be 1-2 kg daily. These symptoms are similar to diverse diseases also like it occurs in ageing persons, thus it becomes a complex task to get a correct diagnosis, which results in death in the near future. In several works, computer technologies might be used for doing accurate diagnoses of patients and distinguish this disease to stop it from becoming deadly. Machine learning and artificial intelligence are playing an enormous role in the medical industry, thus this research work aims to design and develop efficient classification methods for heart disease diagnosis.

1.11 Thesis Organization

In this section, the association of the research work explaining the medical data classification is explained. Chapter 1 illustrates an introduction to cloud computing, data sharing in cloud infrastructures, and disease prediction using the cloud environment. The background of disease prediction using cloud data reveals the motivation behind establishing an effective data classification strategy. Chapter 2 deliberates different existing techniques that donated to the analysis of disease prediction in the field of medicine for taking intelligent decisions for diagnosis. Chapter 3 illustrates the research methodology and the equivalent assessments of the data classification model using Taylor-BSA-DBN and the performance analysis of the proposed method. Chapter 4 discusses the comparative analysis results of the proposed method for medical data classification. Chapter 5 discusses the conclusion and the future work of the proposed method.

CHAPTER 2

LITERATURE REVIEW

The abundance of medical data necessitates the employment of advanced data analysis methods to extort meaningful information. Using statistical and data mining technologies to better data analysis on enormous data sets has long been a problem for researchers. One of the applications wherein data mining methods are proven to be useful is illness detection. In the last decade, heart disease has become the predominant symptom of death worldwide. Many scientists are employing data mining techniques to aid doctors in the identification of cardiac disease. Data mining algorithms are used to evaluate and extract relevant information and knowledge from large amounts of data to help people make better decisions.

The goal of this chapter is to examine the achievements and prospects in this field by focusing on the major elements of data mining techniques like feature selection, classification, and cloud-based approaches, as well as their uses in integrated clinical research in heart disease. The main features of feature selection with traditional methods for disease diagnosis. The main features of classification include machine learning (ML) and Deep Learning (DL) methods for disease diagnosis. This chapter elaborates the review on the different research related to disease prediction using medical data.

2.1 Feature Selection (FS) Methods

The accuracy of Computer-Aided Diagnosis (CAD) technologies has been improved using Machine Learning (ML). This part evaluates the quality of models

created using machine learning approaches by selecting effective attributes using a variety of attribute selection approaches.

The accuracy of Computer-Aided Diagnosis (CAD) technologies has been improved using Machine Learning (ML). This part evaluates the quality of models created using machine learning approaches by selecting effective attributes using a variety of attribute selection approaches.

2.1.1 Filter methods

For the statlog heart-disease database, Jabbar et al. (2015) employed a Feature Selection with a Chi-squared feature assessor in combination with Random Forest (RF) ML methods to develop a framework for heart-disease diagnosis. The database has 270 cases and shares similar features with the Cleveland database. These researchers utilised Chi Squared using backwards removal, in which they rate the attributes using Chi-Squared Test (CST), then eliminate the least matching element one at a time, building and testing a method at every stage till the model's efficiency enhances. The most suitable methodology they identified was 83.7% accurate.

Haq et al. (2018) created a machine-learning-based diagnosis method for predicting heart disease using a heart illness database. Performance assessment criteria like classification accuracy, specificity, sensitivity, Matthews' correlation coefficient, and time complexity were used to evaluate seven common machine learning techniques, three feature selection techniques, the cross-validation approach, and seven classification methods. The proposed technology can quickly distinguish between those with heart disease and healthy people. Receiver optimistic curves and the area under the curves were also calculated for every classifier. The study employs all the classifiers, feature selection techniques, preprocessing methods, validation

methods, and classifier performance evaluation measures. The suggested system's efficiency has been verified using both full features and a limited set of features. The decrease of features influences classification models' performance with regard to accuracy and processing time. The suggested machine learning-based decision support technique will help clinicians in accurately diagnosing heart patients.

Saqlain et al. (2019) proposed a feature subset selection procedure for a clinical cardiac disease detection system for the purpose of improving performance. Three strategies for choosing candidate feature subsets are presented in the suggested methodology: (1) mean Fisher score-based feature selection procedure, (2) forward feature selection algorithm and (3) reverse feature selection strategy. The most conclusive subset from the potential feature subsets is selected using a feature subset identification technique. Features are added to feature subsets based on each Fisher score, and the choosing of a feature subset is based on their Matthews' correlation coefficient value and dimensionality. The selected features subset is loaded into a Radial Basis Function (RBF) kernel-based SVM, which produces a binary classification: (1) heart disease patient and (2) healthy control topic. The accuracy, specificity, and sensitivity of the suggested method are tested using four UCI databases: Cleveland, Switzerland, Hungarian, and Single Proton Emission Computed Tomography (SPECTF). The suggested technique's analytical results are displayed in contrast to current methodologies, demonstrating its superior performance. For Cleveland, Hungarian, Switzerland, and SPECTF, it has the accuracy of 81.19%, 84.52%, 92.68%, and 82.7 %, respectively.

To overcome the feature selection problem, Li et al. (2020) introduced the Fast Conditional Mutual Information (FCMIM) feature selection technique. These techniques are employed to improve the accuracy of the classification and lower

the classification model's completion time. In addition, the leave one topic out cross-validation technique was utilised to learn the best practices in model evaluation and hyperparameter adjustment. The performance measurement metrics are employed to evaluate the classifiers' performance. The effectiveness of the classification models was evaluated using chosen features which were determined through feature selection methods. The experimental outcome suggested using a classification model support vector machine to create a high-level intelligent system to detect heart illness. In comparison to earlier proposed approaches, the suggested FCMIM-SVM diagnosing system exhibited better accuracy. Furthermore, this suggested approach can be simply used in the hospital to detect heart issues.

Spencer et al. (2020)experimentally assess the performance of models derived from machine learning techniques by using relevant features chosen by various feature-selection methods. Four commonly used heart disease datasets have been evaluated using principal component analysis, Chi-squared testing, ReliefF and symmetrical uncertainty to create distinctive feature sets. Then, a variety of classification algorithms have been used to create models that are then compared to seek the optimal feature combinations, to improve the correct prediction of heart conditions. Found the benefits of using feature selection vary depending on the machine learning technique used for the heart datasets we consider. However, the best model we created used a combination of Chi-squared feature selection with the BayesNet algorithm and achieved an accuracy of 85.00% on the considered datasets.

Muhammad et al. (2020) proposed irrelevant and noisy data from extracted feature space, four distinct feature selection algorithms are applied and the results of each feature selection algorithm along with classifiers are analyzed. Several performance metrics namely: accuracy, sensitivity, specificity, AUC, F1-score, MCC,

and ROC curve are used to observe the effectiveness and strength of the developed model. The classification rates of the developed system are examined on both full and optimal feature spaces, consequently, the performance of the developed model is boosted in the case of high variated optimal feature space. In addition, P-value and Chi-square are also computed for the ET classifier along with each feature selection technique. It is anticipated that the proposed system will be useful and helpful for the physician to diagnose heart disease accurately and effectively.

Benhar et al. (2020)developed a comparison of three filter feature ranking methods such as ReliefF, Correlation, and Info Gain. Feature ranking methods need to set a threshold (i.e. the percentage of the number of relevant features to be selected) in order to select the final subset of features. Thus, this study aims to investigate if there is a threshold value which is an optimal choice for three different feature ranking methods and four classifiers used for heart disease classification in four heart disease datasets. The used feature ranking methods and selection thresholds resulted in optimal classification performance for one or more classifiers over small and large heart disease datasets. The size of the dataset takes an important role in the choice of the selection threshold.

Benhar et al. (2020) presented a study aims at evaluating and comparing the performances of six univariate filters: ReliefF, Linear Correlation, Info Gain(IG), Signal-to-noise ratio, minimum Redundancy Maximum Relevance(mRMR), and t-test; and two multivariate filters: Correlation-based feature subset selection and Consistency-based subset selection. The selected features were evaluated with two white-box (K-Nearest Neighbors (K-NN) and Decision Trees(DTs)) and two black-box (Support Vector Machines (SVMs), and Multilayer Perceptron(MLP)) classification techniques using five heart disease datasets. Furthermore, this study

deals with the setting of the hyperparameters' values of the four classification techniques for each feature subset. This study evaluates 600 variants of classifiers. Results show that white-box classification techniques such as K-NN and DTs can be very competitive with black-box ones when hyperparameters' optimization and feature selection were applied. Moreover, the accuracy results of the best performing white-box classifiers of the present study were compared with those from previous studies. In addition to the interpretability advantage, the constructed techniques showed very promising results in terms of accuracy as well.

2.1.2 Wrapper Methods

Usman et al. (2018) developed two distinct cuckoo-inspired methods for feature selection on various heart disease databases: Cuckoo Search Algorithm (CSA) and Cuckoo Optimization Algorithm (COA). Throughout subset creation, both methods employed the generic filter approach. On all databases, the acquired outcomes demonstrated that CSA outperformed COA in connection with total features and prediction accuracy. Ultimately, when CSA was compared against modern techniques, it was discovered that CSA outperformed all other databases.

Takci (2018) combined Machine Learning (ML) and feature selection (FS) techniques. The goal is to find the optimum ML approach and FS method for predicting heart problems. On Statlog (Heart) database, several machine learning techniques with optimal values and many feature selection approaches were tested and assessed for this goal. The Support Vector Machine (SVM) method with the linear kernel is the optimum ML technique, based on the current scientific results, whereas the reliefF technique is the finest feature selection strategy. This combination provides 84.81% accuracy.

Shuriyaa and Rajendranb (2018) suggested an association rule for forecasting disease based on fuzzy logic, with the complexity determined using Adaptive Neuro-Fuzzy Inference System (ANFIS). To achieve the best results, the influence of various values for essential factors was explored. The accuracy for cardiovascular problems is estimated to be 93.12%.

Gokulnath and Shantharajah (2019) suggested an optimization function depending on the concept of SVM. This subjective function is utilised in the Genetic Algorithm (GA) to determine which features are more important in determining heart disease. The GA–experimental SVM's findings are contrasted with Relief, Correlation-based Feature Selection (CFS), Filtered subset, Info Gain (IG), Consistency subset, Chi-squared, One attribute-based, Filtered attribute, Gain Ratio (GR), and GA techniques. The receiver operating characteristic evaluation is used to assess the SVM classifier's efficiency. The Cleveland heart disease database is used to analyze the suggested model in MATLAB simulation.

Gokulnat and Shantharajah (2019) employed a GA to identify characteristics from the Cleveland database. This strategy yielded a subset of seven attributes for which they used four machine learning approaches to develop algorithms for heart disease diagnosis: Support Vector Machine (SVM), Multilayer Perceptron (MLP), J48, and K Nearest Neighbours (KNN). They utilised 10-fold cross-validation to test their algorithms and contrasted these results to models generated utilising actual feature sets, as well as feature sets, chosen by using frequently employed feature selection approaches. The GA attained the best accuracy of 88.34% when combined with SVM, contrasted to 83.70 % with the actual database.

Shah et al. (2020) proposed an automatic investigative methodology for a clinical heart problem. The proposed approach computes the most relevant feature subset by taking advantage of feature assortment and extortion approaches. To accomplish the feature selection, two methods (Mean Fisher-based Feature Selection Algorithm (MFFSA) and Accuracy Based Feature Selection Algorithm (AFSA)). The selected feature subset is additionally distinguished through the feature extraction technique i.e., principal component analysis. The proposed technique is validated over Cleveland, Hungarian, Switzerland and a combination of all of them. To categorize a human as a Heart Disease Patient (HDP) or a Normal Control Subject (NCS), Radial Basis Function Kernel-based Support Vector Machines (RBF-KSVMs) are employed. The suggested technique is assessed by accuracy, specificity and sensitivity metrics.

2.1.3 Embedded method

Reddy et al. (2019) forecasted the classifier and determined which selected features to play a major role in the heart disease diagnosis by employing Cleveland and statlog project heart database. Depending on 3 distinct proportion splits, the accuracy of the RF method in both categorization and feature selection methods was found to be 90–95%. The eight and six chosen features appear to be the bare minimum for creating a superior performance model. Additional lowering of the 8 or 6 chosen features, on the other hand, may not improve the forecast model's performance.

Hogo (2020) studied the patient’s gender effects on the CAD diagnosis model structure and performance. It built two separate and individual models: male and female. The feature set of each model was selected using the Features Ranking Voting (FRV) Algorithm. The memberships of the selected features for each model were computed using the probabilistic clustering technique. Thirty-eight different classifiers

for each model are introduced to select the best one with high performance and a simple structure. The results of each selected diagnosis model of each gender were analyzed and compared with related works. The comparison indicates the suggested technique performs better than present systems and with a simple structure. The high-performance results prove the success of the suggested gender-based technique for the identification of coronary artery illness.

Yadav and Pal (2020) proposed three features-based algorithms: Pearson Correlation, Recursive Features Elimination (RFE) and Lasso Regularization. The data table is analyzed by different feature selection methods for better prediction. All the analysis is done by three experimental setups; the First experiment applied Pearson Correlation on M5P, Random Tree (RT), Reduced Error Pruning (REP) and Random Forest (RF) ensemble method. In the second experiment, RFE is applied to the above four tree-based algorithms. The third experiment used Lasso Regularization and applied as above tree-based algorithms. After all, the performance was analyzed and calculated classification accuracy, precision and sensitivity. With the results finally concluded that feature selection methods Pearson correlation and Lasso Regularization with RF ensemble method provide better results with 99% accuracy.

Zhang et al (2021) proposed a novel heart disease prediction model. A heart disease prediction algorithm is proposed that combines the embedded feature selection method and deep neural networks. This embedded feature selection method is based on the Linear Support Vector Classifier (LinearSVC) algorithm, using the L1 norm as a penalty item to choose a subset of features significantly associated with heart disease. These features are fed into the deep neural network built. The weight of the network is initialized with the initializer to prevent gradient varnishing or explosion so that the predictor can have a better performance. The proposed model is tested on the heart

disease dataset obtained from Kaggle. Some indicators including accuracy, recall, precision, and F1-score are calculated to evaluate the predictor, and the results show that the model achieves 98.56%, 99.35%, 97.84%, and 0.983, respectively, and the average AUC score of the model reaches 0.983, confirming that the method proposed is efficient and reliable for predicting heart disease.

Liu et al (2019) proposed an embedded feature selection method using the proposed weighted Gini index (WGI). Its comparison results with Chi2, F-statistic and Gini index feature selection methods show that F-statistic and Chi2 reach the best performance when only a few features are selected. As the number of selected features increases, the proposed method has the highest probability of achieving the best performance. The area under a receiver operating characteristic curve (ROC AUC) and F-measure are used as evaluation criteria. Experimental results with two datasets show that ROC AUC performance can be high, even if only a few features are selected and used, and only changes slightly as more and more features are selected. The results are helpful for practitioners to select a proper feature selection method when facing a practical problem.

Dissanayake and Md Johar (2021)conducted an experimental evaluation of the performance of models created using classification algorithms and relevant features selected using various feature selection approaches. For results of the exploratory analysis, ten feature selection techniques, i.e., ANOVA, Chi-square, mutual information, ReliefF, forward feature selection, backward feature selection, exhaustive feature selection, RFE, Lasso regression, and Ridge regression, and six classification approaches, i.e., DT, RF, SVM, K-NN, logistic regression, and Gaussian naive Bayes, have been applied to Cleveland heart disease dataset. The feature subset selected by the backward feature selection technique has achieved the highest classification

accuracy of 88.52%, precision of 91.30%, the sensitivity of 80.76%, and f-measure of 85.71% with the DT classifier.

2.1.4 Fuzzy clustering-based feature selection methods

Liu et al (2022) proposed two incremental fuzzy clustering algorithms based on feature reduction. The first uses the Weighted Feature Reduction Fuzzy C-Means (WFRFCM) clustering algorithm to process each chunk in turn and combines the clustering results of the previous chunk into the latter chunk for common calculation. The second uses the WFRFCM algorithm for each chunk to cluster at the same time, and the clustering results of each chunk are combined and calculated again. In order to investigate the clustering performance of these two algorithms, six datasets were selected for comparative experiments. Experimental results showed that these two algorithms could select high-quality features based on feature reduction and process large-scale data by introducing the incremental strategy. The combination of the two phases cannot only ensure clustering efficiency but also keep higher clustering accuracy.

Gu et al (2017) introduced the popular sparse representation method into the classical fuzzy c-means clustering algorithm and presents a novel fuzzy clustering algorithm, called fuzzy double c-means based on sparse self-representation (FDCM_SSR). The major characteristic of FDCM_SSR is that it can simultaneously address two datasets with different dimensions, and has two kinds of corresponding cluster centres. The first one is the basic feature set that represents the basic physical property of each sample itself. The second one is learned from the basic feature set by solving a spare self-representation model, referred to as the discriminant feature set, which reflects the global structure of the sample set. The spare self-representation

model employs the dataset itself as the dictionary of sparse representation. It has good category distinguishing ability, noise robustness, and data-adaptiveness, which enhance the clustering and generalization performance of FDCM_SSR. Experiments on different datasets and images show that FDCM_SSR is more competitive than other state-of-the-art fuzzy clustering algorithms.

Maheshwari and Sharma (2018) present a feature selection strategy to optimize fuzzy-based clustering over very large data. Using the selective features method can reduce the number of features used to classify the following dataset; thus, the reduction of dimensions/features can help in optimizing iteration count, space, and time as well as minimize objective function for the following dataset. In the final observation, found the reduction in iteration count as well as time in comparison to the literal FCM algorithm that is 10.65 s and 9.84 s, respectively, for pen digits and cement dataset.

Ahmadi and Khamforoosh(2020)proposed Firefly Metaheuristic Algorithm (FMA) selects the features through other dataset features at each stage. Features data are clustered using C-means fuzzy clustering to determine clustering accuracy amounts such as Root-Mean-Square Error (RMSE) to specify how useful these features are and how much these selected features have been able to make classified correctly using clustering based on the dataset as well. Regarding this, the target class is predicted according to the selected features, where the results show the optimal performance of the proposed method. Because of using the combination of FMA and FCM clustering, the optimal centres of each cluster are found quickly, the selected feature sets known as the target class representative have the least error value, and the relationship between features is considered as well by completing the iteration of the algorithm.

Souri et al (2017) proposed a new approach for unsupervised feature selection based on the Genetic Algorithm as a heuristic search approach and combine it with the Fuzzy C-Means algorithm. A dual, multi-objective fitness function is introduced based on Davies-Bouldin (DB) and Calinski-Harabasz (CH) indexes. Results show that these indices do not necessarily have similar behaviours. Thus, rather than simply considering their weighted average as a new fitness function, a new approach is proposed to aggregate them based on their tradeoffs. A comparison of the proposed approach with popular feature selection algorithms, across different datasets, indicates the outperformance of the proposed approach for feature selection.

From the review, it concludes that the FS with filter methods gives improved accuracy. Thus filter-based feature selection techniques are employed in this work for pre-processed data before creating the forecasting model by classifiers. Table 2.1 reviews the details of feature selection methods related to the healthcare datasets with their issues and advantages.