Development of ensemble model for heart disease diagnosis

Awotimehin Olasunkanmi Julius

Master Thesis

2023

Tomas Bata University in Zlín
Faculty of Applied Informatics

Tomas Bata University in Zlín
Faculty of Applied Informatics
Department of Informatics and Artificial Intelligence

Academic year: 2022/2023

ASSIGNMENT OF DIPLOMA THESIS

(project, art work, art performance)

Name and surname:	**Olasunkanmi Julius Awotimehin**
Personal number:	**A19889**
Study programme:	**N3902 Engineering Informatics**
Branch:	**Information Technologies**
Type of Study:	**Full-time**
Work topic:	**Ensemble model pro diagnostiku srdečních chorob**
Work topic in English:	**Development of Ensemble Model for Heart Disease Diagnosis**

Theses guidelines

1. Prepare a literature survey on the topic.
2. Select methods suitable for creating an ensemble model.
3. Select the appropriate dataset for the job.
4. Implement the proposed algorithm (model) for heart disease diagnosis.
5. Perform testing, appropriate interpretation of results, and discuss the results obtained.

Form processing of diploma thesis: **printed/electronic**

Recommended resources:

1. ALPAYDIN, Ethem. *Introduction to machine learning*. Third edition. Cambridge, Massachusetts: The MIT Press, [2014], 1 online zdroj (xxii, 613 pages). Adaptive computation and machine learning. ISBN 9780262325745. Dostupné také z: https://proxy.k.utb.cz/login?url=http://ieeexplore.ieee.org/servlet/opac?bknumber=6895440
2. GOODFELLOW, Ian, Yoshua BENGIO a Aaron COURVILLE. *Deep learning*. Cambridge, Massachussetts: The MIT Press, [2016], xxii, 775 s. Adaptive computation and machine learning. ISBN 9780262035613.
3. WITTEN, I. H., Eibe FRANK, Mark A. HALL a Christopher J. PAL. *Data mining: practical machine learning tools and techniques*. Fourth edition. Amsterdam: Elsevier, [2017], xxxii, 621 s. ISBN 9780128042915.
4. *Data science & big data analytics: discovering, analyzing, visualizing and presenting data*. Indianapolis: Wiley, [2015], xviii, 410 s. ISBN 9781118876138.
5. GRUS, Joel. *Data science from scratch*. Sebastopol: O'Reilly, 2015, xvi, 311 s. ISBN 9781491901427.
6. OJEDA, Tony, Sean Patrick MURPHY, Benjamin BENGFORT a Abhijit DASGUPTA. *Practical data science cookbook: 89 hands-on recipes to help you complete real-world data science projects in R and Python*. Birmingham: Packt Publishing, 2014, 380 s. ISBN 9781783980246.
7. MILES, Matthew B., A. M. HUBERMAN a Johnny SALDAÑA. *Qualitative data analysis: a methods sourcebook*. Fourth edition. Los Angeles: SAGE, [2020], xxi, 380 s. ISBN 9781544371856.
8. DORSEY, Richard. *Data analytics*. [CreateSpace Independent Publishing Platform], [2017], 67 s. ISBN 9781547089291.

Supervisors of diploma thesis: **doc. Ing. Roman Šenkeřík, Ph.D.**
Department of Informatics and Artificial Intelligence

Date of assignment of diploma thesis: **December 2, 2022**
Submission deadline of diploma thesis: **May 26, 2023**

Tomas Bata University in Zlín
Faculty of Applied Informatics

L.S.

doc. Ing. Jiří Vojtěšek, Ph.D.
Dean

prof. Mgr. Roman Jašek, Ph.D., DBA
Head of Department

In Zlín December 7, 2022

Name of the student:

Thesis topic:

I hereby declare that:

- I understand that by submitting my Master´s thesis, I agree to the publication of my work according to Law No. 111/1998, Coll., On Universities and on changes and amendments to other acts (e.g. the Universities Act), as amended by subsequent legislation, without regard to the results of the defence of the thesis.
- I understand that my Master´s Thesis will be stored electronically in the university information system and be made available for on-site inspection, and that a copy of the Master´s Thesis will be stored in the Reference Library of the Faculty of Applied Informatics, Tomas Bata University in Zlín, and that a copy shall be deposited with my Supervisor.
- I am aware of the fact that my Master´s Thesis is fully covered by Act No. 121/2000 Coll. On Copyright, and Rights Related to Copyright, as amended by some other laws (e.g. the Copyright Act), as amended by subsequent legislation; and especially, by §35, Para. 3.
- I understand that, according to §60, Para. 1 of the Copyright Act, TBU in Zlín has the right to conclude licensing agreements relating to the use of scholastic work within the full extent of §12, Para. 4, of the Copyright Act.
- I understand that, according to §60, Para. 2, and Para. 3, of the Copyright Act, I may use my work - Master´s Thesis, or grant a license for its use, only if permitted by the licensing agreement concluded between myself and Tomas Bata University in Zlín with a view to the fact that Tomas Bata University in Zlín must be compensated for any reasonable contribution to covering such expenses/costs as invested by them in the creation of the thesis (up until the full actual amount) shall also be a subject of this licensing agreement.
- I understand that, should the elaboration of the Master´s Thesis include the use of software provided by Tomas Bata University in Zlín or other such entities strictly for study and research purposes (i.e. only for non-commercial use), the results of my Master's Thesis cannot be used for commercial purposes.
- I understand that, if the output of my Master´s Thesis is any software product(s), this/these shall equally be considered as part of the thesis, as well as any source codes, or files from which the project is composed. Not submitting any part of this/these component(s) may be a reason for the non-defence of my thesis.

I herewith declare that:

- I have worked on my thesis alone and duly cited any literature I have used. In the case of the publication of the results of my thesis, I shall be listed as co-author.
- That the submitted version of the thesis and its electronic version uploaded to IS/STAG are both identical.

In Zlín; dated: .....

Student´s Signature

ABSTRAKT

Vývoj ensemble modelu pro diagnózu srdečních chorob se v posledních letech stává stále populárnějším přístupem díky schopnosti zvýšit přesnost a robustnost tradičních modelů strojového učení. Tato práce představuje studii vývoje a hodnocení ensemble modelu pro diagnózu srdečních chorob. Navrhovaný model využívá několik algoritmů strojového učení, včetně KNN, logistické regrese, vícevrstvého perceptronu (MLP_ANN) a metody podpůrných vektorů, tak aby spojil vlastnosti každého jednotlivého algoritmu a dosáhl přesnějšího výsledky. Výběr features zahrnuje metody chi-kvadrát a informační zisk a výkonnost modelu je hodnocena pomocí standardních metrik. Výsledky ukazují, že ensemble model dosahuje lepších výsledků než tradiční jednotlivé modely a dosahuje vyšší úrovně přesnosti při diagnostice srdečních chorob. Ensemble model také projevuje vylepšenou robustnost, což je klíčové pro aplikace v reálném světě. Tato práce tak poskytuje cenné poznatky pro možný vývoj efektivnějších nástrojů pro diagnostiku srdečních chorob v budoucnosti.

Klíčová slova:

Ensemble model, Diagnóza srdečních chorob, Algoritmy strojového učení, KNN, Logistická regrese, Vícevrstvý perceptron (MLP_ANN), Metoda podpůrných vektorů, Feature selekce, Chi-kvadrát, Informační zisk.

ABSTRACT

Development of an ensemble model for diagnosing heart diseases has become increasingly popular in recent years due to its ability to enhance the accuracy and robustness of traditional machine learning models. This study presents the development and evaluation of an ensemble model for diagnosing heart diseases. The proposed model utilizes several machine learning algorithms, including KNN, logistic regression, multilayer perceptron (MLP_ANN), and support vector machines, to combine the strengths of each algorithm and achieve more accurate results. Feature selection methods, such as chi-square and information gain, are employed, and the performance of the model is evaluated using standard metrics. The results demonstrate that the ensemble model outperforms individual traditional models and achieves higher accuracy in diagnosing heart diseases. The ensemble model also exhibits improved robustness, which is crucial for real-world applications. This work provides valuable insights for the potential development of more efficient tools for diagnosing heart diseases in the future.

Keywords:

Ensemble model, Diagnosing heart diseases, Machine learning algorithms, KNN, Logistic regression, Multilayer perceptron (MLP_ANN), Support vector machines, Feature selection, Chi-square, Information gain.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my supervisor, Prof. Ing. Roman Šenkeřík, Ph.D., for his unwavering support throughout the development of this thesis. His guidance, feedback, and encouragement have been invaluable to me.

I am also grateful to my teachers, Ing. Adam Viktorin, Ph.D., doc. Ing. Zuzana Komínková Oplatková, Ph.D., and Ing. Tomáš Kadavý, for their insightful courses and class work, which laid the foundation for this thesis. Special thanks also go to Prof. Marek Kubalcik, Ph.D., and Ing. Peter Janků, Ph.D., for their support.

I would like to extend my gratitude to the following individuals: Mirka Viragova, Nika Urbanova, Katerina, Gebauer a Griller company, Lepsi prace, and GG staff members in Mikulova (Anicka simonikova, Jana opavska, zaneta Moravcikova) for providing me with the opportunity to work while studying and for their valuable advice and support.

To my classmates and friends Kenedy, Pascal, George, Li Peng, Tola, Solomon, keneath, and Evka, thank you for your support and encouragement.

I am also deeply indebted to my family members, Mr. Awotimehin Dayo, Mr. Balogun Niyi, my mother, Isaac Awo, and Daddy Sammy, for their unwavering support and encouragement.

Finally, I would like to express my gratitude to the Creator of the universe for giving me the strength and determination to complete this thesis.

CONTENTS

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Heart disease is a leading cause of death worldwide, with more than half of the deaths occurring in men. In the United States alone, over 610,000 people die annually due to heart disease. The Czech Republic has a high mortality rate of 1,077 people per 100,000 due to cardiovascular diseases and strokes, with coronary heart diseases responsible for 34.04% of overall deaths in the country according to WHO data from 2020 [87] and Czech Society of Cardiology[1].

Cardiovascular diseases are the leading cause of death worldwide, with projections suggesting that they will continue to be a major health concern in the years to come, especially with the aging population. Coronary heart disease (CHD) is a highly lethal disease that occurs when there is a build-up of blockage in the coronary arteries, leading to a reduction in blood flow to the heart and an increased risk of heart attack or stroke. High blood pressure, high cholesterol, and smoking are three key risk factors for heart disease [97] and several other medical conditions and lifestyle choices, including diabetes, obesity, poor diet, physical inactivity, and excessive alcohol consumption, can also increase the risk of heart disease.

Despite the availability of treatment options for heart disease, the burden remains high, particularly with cardiovascular diseases, due to the inability to match patients to treatments that are most suitable for them individually.

Soft computing tools can aid physicians in making prompt and accurate decisions. Therefore, data analytics, a modern and popular field of research, can assist doctors and medical professionals in making better and timely decisions by leveraging data mining.

Medical research and discovery are advancing rapidly, making computer-aided diagnosis more intelligent and desirable, and an invaluable instrument in the healthcare industry. Recent developments in Artificial Intelligence (AI) provide techniques that can potentially solve previously difficult tasks in the medical field using computer-based systems. Research is underway globally to explore the new applications of AI in medicine, particularly in diagnosis [33].

[1] https://www.kardio-cz.cz/data/clanek/902/dokumenty/cardiovascular-prevention-in-czech-republic.pdf

New studies reveal that healthcare personnel make predictions on a daily basis. Using this new method, they classify patients based on their condition, offer prognoses of their future health and well-being, and make classifications based on laboratory results. With the introduction of electronic and well-integrated hospital information management systems, the ability to display and use computerized predictive models for tasks like these is significant.

The introduction and emergence of Information Technology (IT) has thrown it opened for unprecedented opportunities and benefit in health care delivery system as the demand for artificial intelligent and knowledge-based system has tremendously increased as modern medical practices and obligations become more knowledge-intensive [35]. Computer results are 99.9% accurate and is far better than humans in the ability to remember and document things and such as characteristic is very sufficient and valuable for a computer aided system that enables more improvements in both diagnosis and treatment [50].

Machine learning and AI can uncover hidden patterns in medical datasets, which can be valuable for clinical diagnosis [52].

Machine learning methods and techniques are often being used in the modern days especially in medical diagnosis. The use of different signs and symptoms for the diagnosis of disease does not actually mean that other diagnostic tools are not available [32]. The problem now is that they are very few well trained or qualified medical personnel most especially in the rural, village or small geographical areas to interpret test results. The main objective and primary goals of machine learning is to forecast and predict an “unknown” value of a newly acquired sample from observed samples, such as forecasting which is achieved or gotten by two sequential phases: (a) Training phase – This produces a predictive model from training samples using one of the available supervised learning algorithms; and (b) testing phase: evaluating all the general predictive model using some testing samples that are not used in the training phase.

1.2 RESEARCH MOTIVATION

A thorough review of the latest research and literature on heart disease shows that coronary artery disease (CAD) remains one of the leading causes of death worldwide. According to WHO (2015), an estimated 7.2 million cases were due to coronary artery disease, the most common type of heart disease. Low- and middle-income countries are disproportionately

affected, with 82% of heart disease deaths occurring in low- and middle-income countries, with a roughly even distribution between men and women [87].

By 2030, approximately 23.6 million people worldwide are expected to die from heart disease, mainly from coronary artery disease and stroke. They are expected to remain the single leading cause of death. "Therefore, more prudent and efficient cardiac therapy and regular screening is of great importance [88]."Therefore, the thesis concluded that further efforts are needed to reduce the large number of CAD-related deaths. Expert medical advice is a scarce, costly, but important component of the healthcare system. Although the world's population is growing, the growing number of cardiologists is inadequate or even in short supply in many places where cardiologists are needed most.

Due to the adverse effects of the CHD in both developed and developing countries, researchers in the AI field have conducted a lot of research to develop systems that can diagnose cases of CHD. Several works on a heart disease diagnosis decision support system using neural networks for prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, and Enhanced Prediction of Heart Disease through Genetic Algorithm and RBF Network [47, 5].

Several research works have highlighted the importance of further research in the field of coronary heart disease. The research include the development of decision guide machines using manual vector machines and artificial neural networks [6], the utilization of genetic algorithms and neural networks for prediction systems and the application of conventional and neuro-fuzzy frameworks for diagnosis and treatment [50], the use of neural network approaches for diagnosing heart disease and the development of fuzzy logic-driven expert systems for diagnosing heart failure disease [2],the effective prediction of heart disease using data mining techniques and the diagnosis of coronary heart disease using ensemble machine learning [32].

Other research emphasize the importance of early detection and intervention in reducing the severity of heart disease. By leveraging computing technology and machine learning tools, physicians can improve their ability to diagnose and predict the disease, enabling them to provide timely treatment and potentially prevent adverse outcomes, including the possibility of death. Predicting disease progression is one of the most interesting and challenging tasks in developing data mining applications. Accurate automated classification systems in heart

disease screening and classification can help reduce the burden on health care workers in early detection of heart disease [101].

Therefore, the state-of-the-art in data mining science centred on ensemble learning approaches to create models and methods that can assist medical physicians in detecting CHD. Ensemble learning algorithms use traditional machine learning algorithms to generate multiple base models and combine them into an ensemble model. They often perform much better than individual models. Most existing CAD diagnostic systems use a single predictive model and accuracy is a major drawback. The researched literature also indicates a shortage of medical staff and facilities, especially in developing countries. Rapid and accurate diagnosis of CHD is critical to reducing fatalities and reducing the economic impact of the threat.

1.3 RESEARCH OBJECTIVES

(a) to design stack and voting ensemble prediction models for coronary artery disease

(b) Implement the ensemble model designed in (a) above.

(c) Evaluation of ensemble model performance based on standard metrics.

1.4 DEFINITION OF TERMS

(a) Coronary artery disease (CAD): CAD involves decreased blood flow to the heart muscle due to plaque building up in the arteries of the heart.

(b) Record/dataset: A collection of related data or information composed of discrete elements that can be processed as a unit by a computer.

(c) Diagnosis: The art or practice of identifying disease by signs and symptoms.

(d) Ensemble method: A machine learning technique that combines multiple base models to create an optimal predictive model.

(e) Machine learning is an application of artificial intelligence (AI) that gives systems the ability to automatically learn and improve from experience without being explicitly programmed.

(f) Model accuracy: Machine learning model accuracy is a measure used to determine which model best identifies relationships and patterns among variables in a dataset, given input or training data.

(g) Predictive modelling: It is the process of predicting future outcomes or actions based on past and present data.

1.5 THESIS ORGANIZATION

This work is divided into five parts. Part 1 is an introduction, briefly introducing the research work and some of the concepts used. Part 2 is the theory and also describes some related studies and existing systems with similar views to this study. Part 3 contains structural points that describes how solutions to the limitations identified by the review of existing work were systematically formulated and the selection of methods suitable for creating an ensemble model. Part4 focuses on system results, discussion, and the proposed algorithm for heart disease diagnosis. Part 5 contains appropriate interpretation of results and provided conclusions and recommendations, as well as areas for further research.

I. THEORY

2 CORONARY HEART DISEASE

Heart disease is a leading cause of death worldwide, receiving significant attention in medical research. The World Health Organization reports that this disease is responsible for the most fatalities in both developed and developing countries [88].

Some of the risk factors associated with heart disease include age, blood pressure, smoking, cholesterol, diabetes, hypertension, family history of heart disease, being overweight, and lack of proper physical activity. In order to identify patients at high risk of heart disease, it is important to have more understanding of these risk factors [72, 40].

Other forms of heart disease include irregular heartbeat (arrhythmias), congenital heart defects, weak heart muscles (cardiomyopathy), heart valve problems, heart infections, and cardiovascular disease.

Coronary heart disease (CHD) is the most frequent type of heart problem and the leading cause of heart attacks. CHD refers to damage to the heart that occurs because its blood supply is insufficient. A large build-up of fatty deposits on the linings of the blood vessels that provide blood to the heart muscles causes them to narrow, decreasing blood flow and resulting in pain known as angina. A sticky deposit comprising cholesterol and other substances is deposited in the arterial wall. CHD is deadly, accounting for 7 million deaths annually worldwide [56].

The prevalence or reoccurring of CHD is on high increase all over the world. It often results into death and has been categorized as one of the world's most predominant causes of death [104].

Coronary heart disease (CHD) can lead to myocardial deterioration and heart failure over time, with symptoms such as arrhythmia indicating an early stage of the disease. Diagnosing CHD involves considering various factors including blood pressure, chest pain, cholesterol levels, fasting blood glucose, exercise-induced angina, resting electrocardiogram results, high body mass index, and physical activity levels, Age, gender, and family history of heart disease may also contribute to risk factors for CHD[2]. If a patient is symptomatic or at high risk, a cardiologist will use multiple tests to diagnose CHD and prescribe an appropriate treatment plan, though this process can be expensive and resource-intensive. Research done

[2] https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronary-heart-disease

suggests that most people develop plaques associated with CHD during their teenage years, emphasizing the importance of early prevention and lifestyle changes to mitigate the disease's progression. Figure 1 shows the pictorial representation of Coronary Artery disease.

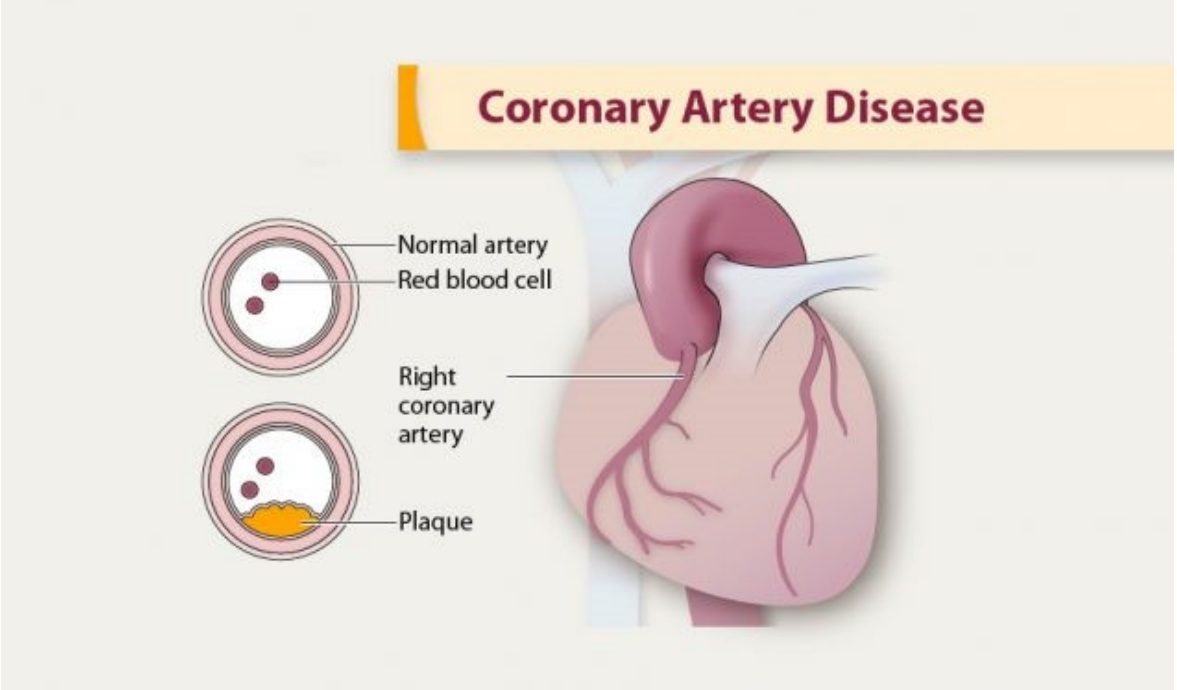


Figure 1 Pictorial Representation of Coronary Heart Disease

2.1 RISK FACTORS OF CORONARY HEART DISEASE

According to the American Heart Association (2016), several risk factors have been identified for the development of coronary heart disease (CHD).

Age is a significant risk factor, with the majority of CHD cases and deaths occurring in individuals aged 65 or older.

Gender is also a factor, with men having a greater risk of heart attack than women, and typically experiencing attacks at a younger age [88].

Family history of heart disease is another risk factor, as individuals with a genetic predisposition to the disease are more likely to develop it themselves. Tobacco smoke, whether through smoking or exposure to second hand smoke, is a powerful independent risk factor for CHD.

Finally, high blood cholesterol levels have also been identified as a significant risk factor for CHD. As blood cholesterol rises, so does the risk of CHD. We have some other risk factors (such as high blood pressure and tobacco smoke) are found in the body, this risk increases even more. A person's cholesterol level is also affected by age, sex, heredity and diet.

High blood pressure

High blood pressure increases the heart's workload, causing the heart muscle to thicken and become stiffer. This stiffening of the heart muscle is not normal, and prevents the heart working properly. It also add more to the risk of stroke, heart attack, kidney failure and congestive heart failure.

Physical Inactivity

An inactive lifestyle is a risk factor for coronary heart disease. Consistent, Regular, moderate-to-vigorous physical activity helps reduce the risk of heart and blood vessel disease. Even moderate-intensity activities are helpful if done regularly and over a long term. Physical activity can help control blood cholesterol, diabetes and obesity, and with some people, lower blood pressure.

Obesity and overweight

People who have excess body fat — especially at the waist — are more likely to develop heart disease and stroke even if no other risk factors are present. Overweight and obese adults with risk factors for cardiovascular disease such as high blood pressure, high cholesterol, or high blood sugar can make lifestyle changes to lose weight and produce clinically meaningful reductions in triglycerides, blood glucose, HbA1c, and the risk of developing Type 2 diabetes [11].

2.2 MACHINE LEARNING APPROACH

Machine learning is a subfield of artificial intelligence that systematically launched and applies algorithms to synthesize the underlying relationships among data and information. It is a special field of computer science that gives computer systems the ability to learn (i.e. progressively improve performance on a specific task) with data, without being explicitly programmed. It also evolved from the study of digital pattern recognition and computational learning theory in artificial intelligence [9].

ML is widely used in various applications such as web search, advertising, credit scoring, stock market forecasting, gene sequencing, and big data analytics. It plays a crucial role in developing user-centric innovations by characterizing underlying relationships within large datasets to solve problems in big data analytics, behavioral pattern recognition, and information evolution [7].

One of the main advantages of ML is its ability to generalize the training experience to predict future data instances. ML systems can also classify changing conditions in processes and model variations in manipulative behavior. Unlike other optimization problems, ML does not have a clearly defined function that can be accelerated or optimized. Instead, the process of collating and generalization requires classifiers that input discrete or continuous feature vectors and output a class.

The main goal of ML is to predict future events or scenarios that are unknown to the computer. Arthur Samuel described ML as the "field of study that gives computers the capability to learn without being explicitly programmed" in 1959 [44]. Tom M. Mitchell's definition of ML states that a computer system is programmed to learn from experience and a measure of performance on a task if performance on the task is improved by experience [44].

Alan Turing's seminal work in computing and artificial intelligence (Turing, 1950) set the benchmark standards for demonstrating machine intelligence. Learning experiments and processes play an important role in generalizing problems based on historical experience, which are captured in the form of training data. Datasets that help you achieve accurate results on new, unknown tasks. A training dataset contains an existing problem domain that learners use to build a generic model about that domain. This allows the model to actually produce very accurate predictions on new cases.

2.3 CONCEPTUAL TERMS IN MACHINE LEARNING

In the domain of machine learning, conceptual notions and it encompass the fundamental ideas, principles, and components that form the basis for understanding and applying machine learning algorithms. These terms serve to define the concepts and methodologies utilized in the field, establishing a shared lexicon for discussing and examining machine learning procedures.

(a) Feature Vector

An n-dimensional numeric vector of explanatory variables representing instances of the object to facilitate processing and statistical analysis. Feature vectors are often weighted to build prediction functions that are used to assess the quality or validity of predictions. The dimensionality of the feature vector can be reduced by various dimensionality reduction techniques such as*:* Principal Component Analysis (PCA), Multilinear Subspace Reduction, Isomap, and Latent Semantic Analysis (LSA).

(b) Dimension

A set of attributes that portrays a property. The main primary functions of dimension are filtering, classification, and grouping.

(c) Instance

An object attributed by feature vectors from which the model is either trained for generalization or used for future prediction.

(d) ***Data Mining***

The extraction and analysing of knowledge discovery or pattern detection in a large dataset. The methods and concepts involved in data mining aid in extracting the accurate data and transforming it to a known structure for further evaluation.

(e) Model

A model is a well-organized structure that defines and summarizes a dataset for description or prediction. Each model can be changed to the specific requirements of an application. Applications with big data will have large datasets and there will be many predictors and features that are too huge for a simple parametric model to extract vital and useful information. The learning process synthesizes the parameters and the topology of a model from a given dataset.

(f) Supervised learning

Learning methodology and techniques that extract groups between independent quality and a designated dependent quality (the label). Supervised learning make use of a training dataset to construct a prediction model by implementing the input data and output values.

The model will now be able to make predictions of the output values for another new dataset.

(g) Unsupervised Learning

Learning techniques that classified instances without a pre-specified dependent feature. This methodology generally involves learning structured signs in the data by ejecting pure unstructured vociferation. Clustering and dimensionality reduction algorithms are usually referred to as unsupervised.

(h) Induction Algorithm

An algorithm that make use of the trained dataset to generate a model that generalizes beyond the training dataset.

(i) Confusion Matrix

The Confusion Matrix, also known as the Error Matrix, provides a comprehensive assessment of a classification algorithm's performance using the data within the matrix. It enables a comparison between the predicted classifications and the actual classifications, presenting information on false positives, true positives, false negatives, and true negatives. The visual representation of confusion matrix can be found in Figure 2.

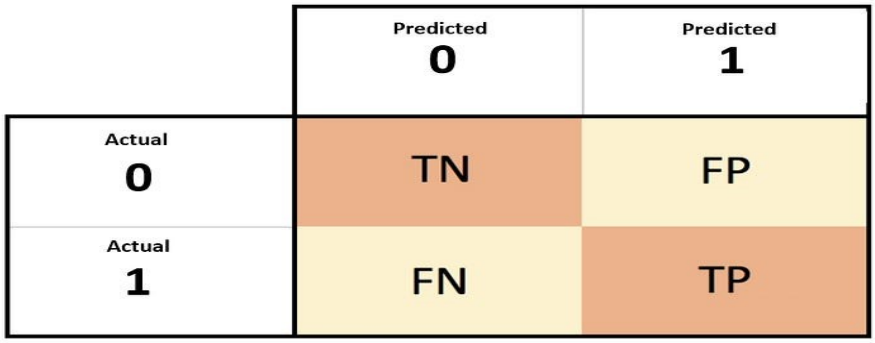


Figure 2 A confusion matrix for a two-class classifier system [3]

(j) Precision (Error Rate)

Accuracy and precision are both important metrics in evaluating the performance of a machine learning model. Accuracy refers to the rate of correct predictions made by the model

[3] https://medium.com/@rahul.apiit23/decode-confusion-matrix-bb554c299d01

over a trained dataset, while precision (also known as error rate) measures the quality or state of satisfaction with the final results of all training. To accurately estimate accuracy, it's common to use an estimated independent test set that was not used during the learning process. More complex techniques like cross-validation and bootstrapping may also be used, particularly for well-trained datasets with a small number of instances.

(k) Cross Validation

Cross-validation is a technique used to evaluate the generalization performance of a model on an independent dataset. It involves partitioning the training dataset into k mutually exclusive folds of equal size, and then training the model *k* times, each time using a different fold for validation while the remaining folds are used for training. The results of the k-fold cross-validation are averaged to estimate the accuracy of the model. This technique is useful in detecting overfitting and in comparing the performance of different prediction functions[4].

2.4 DEVELOPING A MACHINE LEARNING METHOD

Researchers, analysts, and programmers in various fields have developed different theoretical frameworks to understand machine learning methods, including computational learning theory, Bayesian learning theory, classical statistical theory, minimum description length theory, and statistical mechanics approaches [63].

Machine learning aids in the development of programs that improve their performance for a given task through experience and training [90]. Most of the voluminous data applications leverage ML to operate at highest efficiency. The speed, total volume, and diversity of data flow have made it impracticable to exploit the natural capability of human beings to analyse data in real time. The rush in social interacting and the wide use of Internet based applications have resulted not only in greater volume of data, but also increased complexity of data. To preserve data resolution and avoid data loss, these streams of data need to be analysed in real time.

The development of ML starts by identifying all the metrics that are critical to a decision process. The processes of ML synthesize models for optimizing the metrics. Because the

[4] https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

metrics are essential for developing solutions for specific decision-making processes and should be chosen carefully during the conception stage [90]. It is also important to evaluate whether ML, by its very nature, cannot provide perfect accuracy. For solutions that require highly accurate results in a limited amount of time, ML may not be the preferred approach.

Having an excessively high precision is not favourable, and extensive datasets may hold hidden patterns or synthetic information that has yet to be discovered. The problem at hand is not fully comprehended due to a lack of foundational knowledge and historical data, which are essential for devising appropriate algorithms. Additionally, the problem necessitates adaptation to changing environmental circumstances. The process of developing machine learning algorithms can be divided into the following steps, given in next subsections 2.4.1-2.4.7 [90].

2.4.1 Data Collection

Collecting data is an important step in the process of machine learning. It involves gathering relevant information and observations that will be used to train and develop machine learning models. Data collection can be performed through various methods, such as surveys, experiments, observations, or accessing existing databases. The collected data should be representative of the problem domain and cover a wide range of scenarios and instances. Careful consideration should be given to ensure data quality, accuracy, and completeness. Proper data collection lays the foundation for effective analysis, modeling, and decision-making in machine learning applications.

2.4.2 Data Preprocessing

Data pre-processing refers to the steps taken to transform raw data into a format that is suitable for analysis and modeling in machine learning. It involves cleaning the data, handling missing values, dealing with outliers, normalizing or scaling the data, and encoding categorical variables.

Cleaning the data involves removing any irrelevant or unnecessary information, correcting errors, and handling inconsistencies or duplicates. Missing values can be addressed by either removing the corresponding data points, imputing the missing values based on statistical methods, or using advanced imputation techniques.

2.4.3 Data Transformation

Transform algorithm-specific data and problem knowledge. Transformations can take the form of scaling, decomposing, or aggregating features. Features can be decomposed to extract useful components embedded in the data, or multiple instances can be aggregated to combine them into a single feature.

2.4.4 Algorithm Training

Select training and test datasets from the transformed data. The algorithm is trained using the training data set and compared to the test set. The transformed training dataset is fed to algorithms to extract knowledge or information. This trained knowledge or information is saved as a model that is used for cross-validation and practical use. Unsupervised learning without targets does not require a training step.

2.4.5 Algorithm Testing/validation

Evaluate algorithms to test their effectiveness and performance. This step allows you to quickly determine whether you can identify learnable structures in your data. A trained model published to a test dataset is measured against predictions made on that test dataset, indicating the model's performance. If you need to improve the performance of your model, change the data stream, sample rate, transformation, linearization model, outlier removal method, warping scheme, and repeat the previous steps.

2.4.6 Application of Reinforcement Learning

Reinforcement learning is a subfield of machine learning that focuses on training agents to make sequential decisions in an environment. It is commonly applied in various domains where the agent interacts with the environment and learns through trial and error to maximize a reward signal.

One prominent application of reinforcement learning is in the field of robotics. By applying reinforcement learning techniques, robots can learn to navigate and perform complex tasks in dynamic and uncertain environments. For example, a robot can learn to grasp objects, walk, or even perform more sophisticated tasks like playing sports or assisting in medical procedures.

2.4.7 Execution

Apply validated models to perform real-world prediction tasks. When new data is detected, the model is retrained using the previous steps. The training process can coexist with the real task of predicting future behaviour.

2.5 MACHINE LEARNING ALGORITHMS

Based on the basic mapping between the input data presented during the ML learning stage and the expected output, ML algorithms can be classified into six categories:

2.5.1 Supervised Learning Algorithms

It is a learning mechanism that infers underlying relationships between observed data (also called input data) and target variables (dependent variables or labels) on which predictions are based. The learning task uses tagged training data (training examples) to synthesize a model function that attempts to generalize the underlying relationship between the feature vector (input) and the monitor signal (output). Feature vectors influence the direction and magnitude of change to improve the overall performance of the functional model. The training data includes the observed input (feature) vector and desired output values (also called monitor signals or class tags).

Well-trained feature models based on supervised learning algorithms can accurately predict the class labels of hidden phenomena embedded in unknown or unobserved data instances. The goal of any learning algorithm is to minimize the error for a given input set (training set). However, the model can run into overfitting problems when using a low-quality training set that suffers from accuracy and versatility of labelled examples. This usually represents poor generalization and misclassification. Figure 4 showcases the relevant Flow of supervised learning.

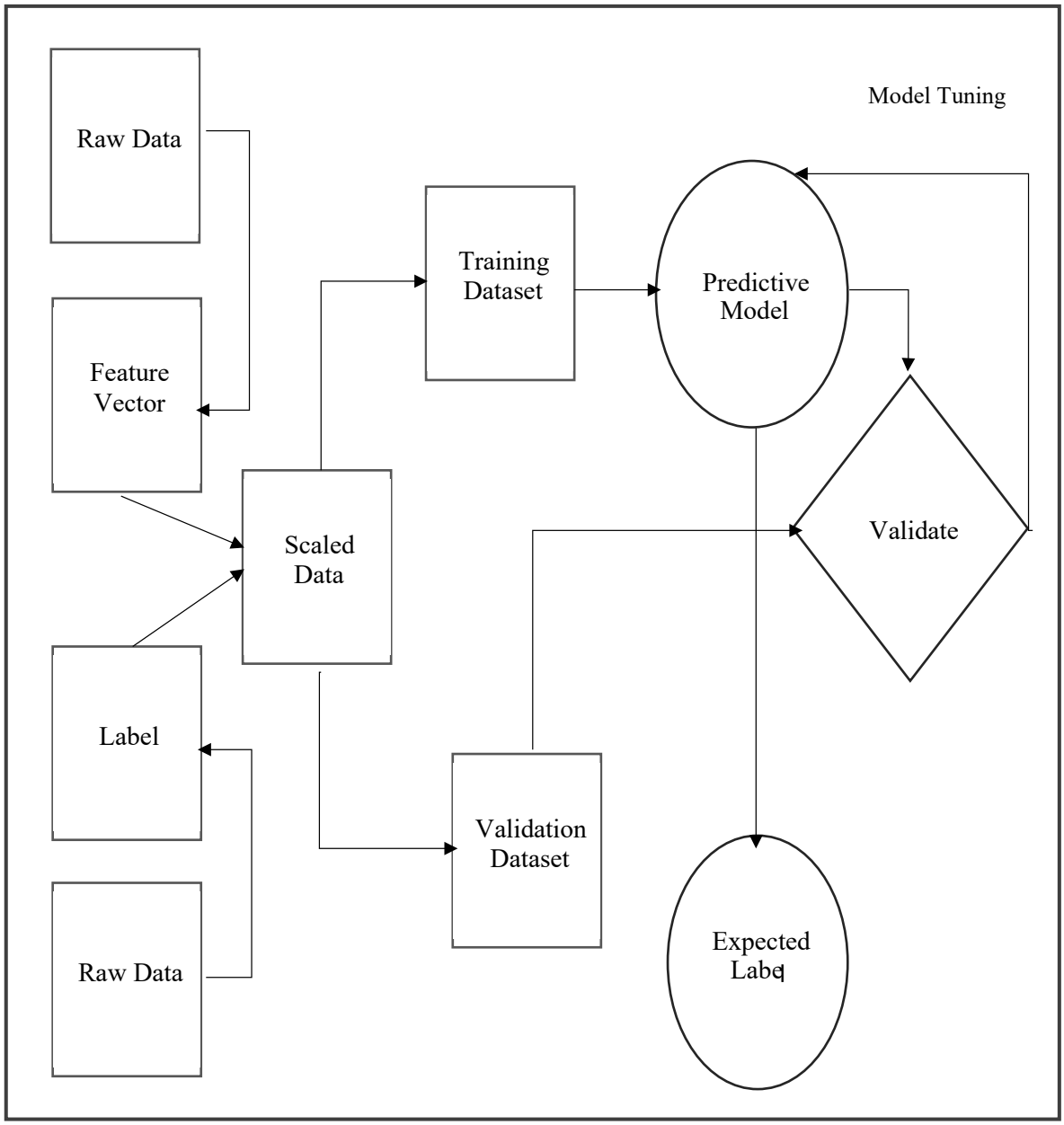


Figure 3 Supervised Learning Flow

The resulting classifier is used to assign class labels to test instances with known predictor values and unknown class labels for those values. The input data for the taxonomy or classification is a set of instances, where each instance is a record of data in the form *(x, y)*. Here, *x* represents the input variables or features, and *y* represents the target variable or class labels. A classification model is a tool that can be used to describe data (descriptive model) or predict the class label of a new instance (predictive model) [12].

In supervised learning, the goal is to learn the associations between a set of input variables (features/attributes) represented as *X* and the output variable represented as *Y*.

The variable j is used to represent the number of records or cases in the dataset. This mapping learned from the training data is then applied to unseen data, which contains the values of the input variables *X* but does not contain the corresponding output variable *Y*.

The aim is to predict the output or class label for the unseen data. Supervised machine learning is widely used in various fields, including engineering and medicine, and it is a popular technique in the field of machine learning.

2.5.2 Unsupervised Learning Algorithms

Unsupervised learning is a mechanism designed to discover hidden structures in unlabelled datasets where the desired output is unknown. This mechanism has many applications in data compression, outlier detection, classification, human learning, and more. Common learning approaches include training with probabilistic data models. Two popular examples of unsupervised learning are clustering and dimensionality reduction. In general, the unsupervised training set consists of inputs $\{x_1, x_2, x_3,..,x_m\}$ but does not include the target output or the reward from its environment (as in supervised learning). The goal of ML in this case is to hypothesize about the representations of input data for efficient decision-making, forecasting, information filtering, and clustering. For example, unsupervised training is useful for developing phase-based models where each phase synthesized through the unsupervised learning process represents a unique condition for opportunistic tuning of the process. In addition, each phase acts as a state and is subject to aggressive resource allocation or allocation forecasting. Unsupervised learning algorithms centred on probability distribution models commonly use Maximum Likelihood Estimation (MLE), Maximum Posterior Probability (MAP), or Bayesian methods. Other algorithms not based on probability distribution models can use statistical measures, quantization error, variance preservation, entropy gaps, and more.

2.5.3 Semi Supervised Learning Algorithm

Semi-supervised machine learning algorithms leverage a combination of labelled and unlabelled datasets to generate model functions or classifiers. While tagged data is essential for training, it is often expensive and impractical to obtain, requiring intensive and skilled human labour. In contrast, unlabelled data is often readily available and relatively cheap. Semi-

supervised learning methodologies occupy a middle ground between unsupervised learning (unlabelled data) and supervised learning (labelled data) and can lead to significant improvements in learning accuracy. Recently, semi-supervised learning has gained importance due to the availability of large amounts of unlabelled data in various applications such as web data, messaging data, inventory data, retail data, biological data, and images [23]. This learning methodology has practical and theoretical value, particularly in areas related to human learning such as language, vision, and handwriting, which require less direct instruction and a greater amount of unlabelled experience.

2.5.4 Reinforcement Learning (RL) Methodology

The field of reinforcement learning involves the study of adaptive sequences of actions or behaviours by intelligent agents in specific environments with the goal of maximizing cumulative rewards. As intelligent agents take actions, observable changes occur in the state of the environment. Learning techniques are used to synthesize adaptive models by training themselves on a given set of experimental actions and observed responses to environmental conditions. This approach can be viewed as a trial-and-error learning paradigm for control theory with rewards and penalties associated with a set of actions. Reinforcement learning agents change policies based on collective experience and resulting rewards, searching for previously investigated actions that led to a reward. However, many unproven actions must be tried in order to build an exhaustive database or model of all possible action-reward predictions. Finding a balance between exploring new potential behaviours and the likelihood of failure resulting from those behaviours is crucial[5] .Key elements of Reinforcement Learning include:

Policies are a key component of the Reinforcement Learning Agent that map control actions to the perceived state of the environment.

Critics represent an estimate function that criticizes actions taken under existing policies. Alternatively, critics rate the performance of the current state according to actions taken under current policy. Criticism agents shape policy by making continuous and continuous revisions.

[5] https://www.sciencedirect.com/science/article/pii/S2667096822000374

The reward function estimates the current desirability of the perceived state of the environment for the attempted control action.

Models are planning and strategy tools that is strongly recommended to aid in predicting the future course of action by contemplating possible uncertain situations.

2.5.5 Transductive Learning (Transductive Inference)

Local models are used to predict the model function for a given test case by incorporating additional observations from the training data set related to new cases [63]. Each new observation is fit to a point in space, resulting in a local model. Unlike global models, new data must fit existing models without making any assumptions about them. In some cases, it may not be necessary for new data to fit the global model completely, and there may be discontinuities during model development. In such situations, multiple models can be created at the boundaries to account for these discontinuities. Newly observed data is processed through the model that satisfies the boundary conditions for which the model is valid.

2.5.6 Inductive Inference

Machine learning algorithms estimate a model function based on the data and use it to predict output values for examples beyond the training set. Some of the functions used include linearly weighted polynomials, logic rules, and Bayesian networks. To reduce the error, statistical learning methods typically start with initial solutions in the hypothesis space and iteratively develop them. Examples of common algorithms that fall into this category include SVM, neural network models, and neuro-fuzzy algorithms. In some cases, lazy learning models can also be applied. The generalization process is an ongoing task that develops a richer hypothesis space based on new data applied to existing models.

2.6 FEATURE SELECTION

Feature selection is a crucial pre-processing step in machine learning, involving the identification and removal of irrelevant or redundant features in a dataset. This step reduces the dimensionality of the data, enabling data mining algorithms to work more efficiently and effectively. Feature selection is considered one of the most important steps in machine learning, as it can significantly improve the accuracy of computer intelligence learning. In fact, different subsets of features can lead to varying performances of the same training data [36].

The effectiveness of machine learning is impacted by a range of factors, with one of the most critical being the quality of the data used in the learning process. Real-world datasets may contain irrelevant or redundant information that can impede the learning process and compromise the accuracy of the results .To address this issue, feature selection techniques can be used to remove such extraneous data during the training phase [30].

Feature selection is a crucial step in data pre-processing, especially in fields like microarray data analysis where irrelevant, redundant features, or noise can be present in data. In sparse data sets, where the number of features exceeds the number of samples, the search space becomes sparsely filled, making it challenging to distinguish between relevant and irrelevant data, thus impacting model accuracy [71] Two approaches are typically used for feature selection: individual assessment and subset assessment. Individual assessment involves assigning feature weights based on their relevance, while subset assessment involves ranking subsets of features based on their overall relevance [53]. In evaluation, candidate feature subsets are usually constructed using search strategy. The general procedure and technique for feature selection has four key steps:

- Subset Generation
- Evaluation of Subset
- Stopping Criteria
- Result Validation

Subset generation is a heuristic search method where each state identifies candidate subsets for evaluation in the search space [36].The subset generation process is influenced by two main factors. Firstly, the next generation determines the starting point for exploration, which impacts the direction of exploration. To determine the search starting point for each state, various methods such as forward, backward, composite, weighted, and random can be considered [11]. Secondly, search organizations are responsible for the feature selection process, which involves using specific strategies such as sequential search, exponential search, or random search [30]. The newly generated subset must be evaluated based on specific criteria. Several criteria have been proposed in the literature to determine the quality of a subset of properties. These criteria can be categorized into two groups based on their dependence on mining algorithms: independent and dependent [30].

Independent criteria assess the goodness of a function set or functions based on intrinsic characteristics of the training data without using mining algorithms. In contrast, dependent criteria include a given feature selection mining algorithm for selecting features based on the

mining algorithm's performance applied to a subset of the selected features. To stop the selection process, stopping criteria must be established. The feature selection process stops at the verification step, which involves comparing different tests to previously determined results or results from competing methods using artificial datasets, real-world datasets, or both. A comparison should be performed to validate the feature selection method. The main steps in the feature selection process can be illustrated by Figure 5.

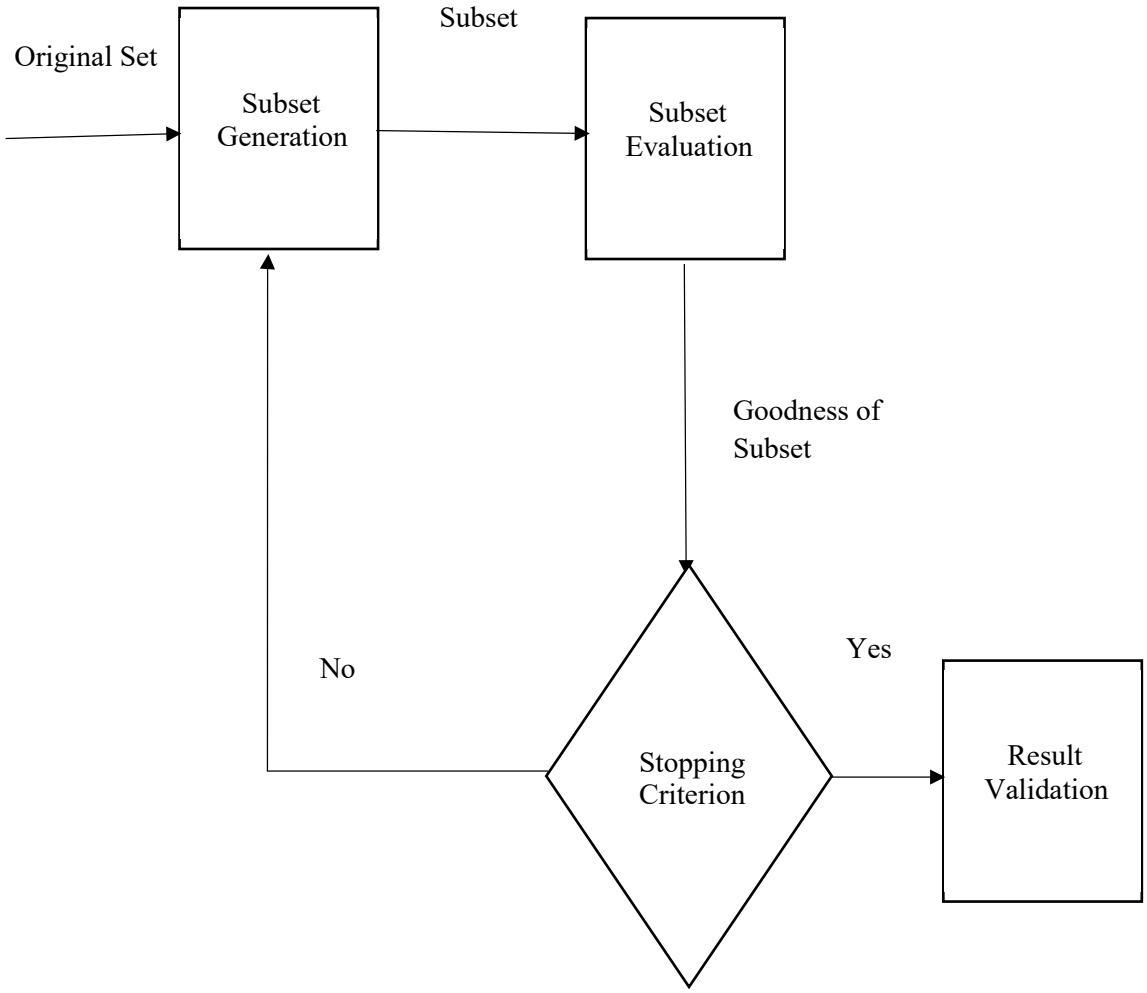


Figure 4 Four critical steps in the feature selection process

The models presented in this thesis are based on the relationship between inductive learning methods and feature selection algorithms. Feature selection can be approached through three general methods. The first is the filtering approach, which exploits general properties of the training data, independent of the mining algorithm [7]. The wrapper approach examines the relationship between relevance and selection of the best feature subset by finding the best

subset of features that fit a given mining algorithm .Finally, the built-in approach involves specific learning algorithms that perform feature selection during the training process.

2.7 DEFINING FEATURE RELEVANCE IN MACHINE LEARNING TASK

The best feature subset consists of the most relevant features. Therefore, it is important to define the relevance of traits accurately. In the literature, features are classified into three categories based on their association: irrelevant, less relevant, and more relevant. In this section, we will discuss the existing definitions of trait relevance proposed in previous studies and also propose degrees of relevance. The dataset S consists of $|S|$ instances, which can be seen as the result of sampling from I . The feature domain $X = \{x_1, x_2, x_3, .., x_m\}$ represents the different features. Let P be the probability distribution of I.

The objective function $C: I \rightarrow L$ assigns labels from the label space L based on the relevance function.

(a) **Definition 1: (Relevance to the target concept)** [30]. A feature *xi* is considered relevant to a target concept c if there are two examples, *A and B,* in the instance space that differ only in their assignment to *xi* and have different classifications according to *c(A) and c(B).*

Alternatively, we can state this definition as follows: a feature *xi* is deemed relevant if there exists an example in the instance space for which changing the value of *xi* influences the classification determined by the target concept.

It is important to note that this definition has a limitation: when the learning algorithm only has access to the sample *S*, it may not be able to definitively determine whether a given feature *xi* is relevant or not.

(b) Definition II (Strong Relevance to the Sample/Distribution) [30]. A feature X_i is considered strongly relevant to a sample *S* if there are examples *A and B* in *S* that differ solely in their assignment to X_i and have distinct labels (or exhibit distinct label distributions if they appear multiple times in *S*). Similarly, x_i is strongly relevant to the target concept c and the distribution *D* if there are examples *A* and *B* that have a non-zero probability over *D* and differ only in their assignment to x_i, satisfying $c(A) \neq c(B)$.

(c) Definition III (Weak Relevance to the Sample/Distribution). A feature X_i is considered weakly relevant to a sample S (or to the target concept c and distribution D) if it is possible to eliminate a subset of features in such a way that x_i becomes strongly relevant.

These notions of relevance provide valuable insights for a learning algorithm in determining which features to retain and which to disregard. Features that are strongly relevant are generally crucial to retain, as removing them introduces ambiguity to the sample. On the other hand, features that are weakly relevant may or may not be important to retain, depending on which other features are discarded [4].

(d) Definition IV (Relevance as a Measure of Complexity Given a data sample S and a set of concepts C, let $r(S, C)$ represent the number of features that are relevant, based on Definition 1, to a concept in C that achieves the lowest error over S and requires the smallest set of relevant features.

In other words, the most interesting point is determining the minimum number of features necessary to achieve optimal performance on S using a concept from C. It is important to specify the concept class C because there may be features, such as a person's social security number, that contain significant information but are irrelevant to the concepts being considered. To enhance robustness, this definition is sometimes adjusted to include concepts in C that have "nearly" minimal error over S if it results in a smaller set of relevant features.

(e) Definition V (Incremental Utility). In the context of a data sample S and a learning algorithm L, a feature X_i is said to be incrementally useful to L with respect to a feature set A if the inclusion of X_i in the feature set $\{X_i\} \cup A$ leads to an improvement in the accuracy of the hypothesis generated by L compared to using the feature set A alone.

(f) Definition VI: Entropic Relevance

This is denoted by mutual information, given $by\ (x; y) = (x) - H(x|y)$, where $H(x)$ represents the Shannon entropy of x. The relevance of x *to* y, denoted as $(x; y)$, is defined as $(x; y)$. Let X be the original set of features and C be the target concept.

The objective can be seen as finding a feature set, $X' \subset X'$ that is sufficient, meaning that X' preserves the learning information. A subset X' is considered more sufficient if it satisfies $(X'; C) = (X, C')$. Therefore, $(X'; C)$ and $(C; X')$ are jointly maximized [80].

(g) Definition VII: Feature Selection

Feature selection is the process of selecting relevant features or candidate subsets of features. Scoring criteria are used to obtain the best feature subset. In high-dimensional data (number of samples < number of features), finding an optimal subset of features is a difficult task. Many related problems have been presented as NP-hard. For data with *N* features, there are *2^N* candidate subsets of features.

Let's assume that the original set of features *A and L (.)* be an evaluation criterion to be maximized (optimized) and well defined as *L: L: A′ ⊆ A* → ℝ. The candidate subset of attributes can be seen and considered under the following considerations [104].

Let |*A*| = m and |*A′*| = *n,* then *(A′)* is maximized, where *m* > *n* and *A′* ⊂ *A*.

Therefore, set a threshold *θ* such that *(A′)* > *θ* to find a subset of the feature along with the smallest number *m* > *n*.

To find the optimization function *(A′)* with optimal feature subsets |*A′*|.

There will be a continuous feature selection problem in which each feature *ak* ∈ *A* is assigned weights *wk* to store the theoretical relevance of the features. Binary weight assignment is considered to be under the binary feature selection problem [14]. The optimal feature subset is considered to be one of the most optimal subsets; therefore, the above definition does not state or ensure that the optimal feature subset is unique. The optimal feature subset is well-defined in terms of induced classifier accuracy as described below.

(h) Definition VIII: Optimal Feature Subset

Let dataset D be given by features *{A₁, A₂, A₃... Aₖ}* from a distribution *P* over the labelled object or instance space and inducer *L*. An optimal feature subset *A_opt* is a subset of the attributes such that the accuracy of the induced classifier *C = L (D)* is maximal [69].

Feature selection can be divided into four main steps: subset generation, subset evaluation, stopping criterion, and result validation. Subset generation involves a search procedure that utilizes a specific search strategy [36]. After generating a subset feature, it is evaluated based on a specific evaluation criterion, and the newly generated subset is compared with the previous best-known feature subset. If the new feature subset is better than the previous one, it replaces it as the previous best feature subset. This process continues until certain stopping criteria are met. Once the stopping criteria are met, the best subset of features generated

needs to be validated. This validation can be done using either synthetic data or real-world datasets [36]. A general feature selection pseudocode:

```
# Inputs
X : Set of features of a data set having n features
SG : Successor Generator Operator
E : Evaluation measure (dependent or independent)
θ : Stopping Criteria
# Output
Xopt : Optimal feature set or weighted features
# Initialize
X_prime = Start point(X)
X_opt = Best of X_prime using E
# Repeat
while True:
   X_prime = Search Strategy(X_prime, SG(E), X)
   X_opt_new = Best of X_prime according to E
   if E(X_prime) >= E(X_opt) or (E(X_prime) == E(X_opt) and len(X_prime) < len(X_opt)):
      X_opt = X_prime
   if Stop Criteria is found:
      break
# Output
Xopt = X_opt
```

2.8 Features selection process

The selection of feature process is a crucial step in machine learning and data analysis. It involves identifying and selecting a subset of relevant features from the original set of features in a dataset.

2.8.1 Subset Generation

Subset generation can simply be illustrated as the process of the heuristic search. The process of subset generation can be divided into two basic issues to determine a feature subset, namely search organization and also successor generation.

(i) Search Organization determines how the search space of possible feature subsets is explored to identify the most relevant features.

For a data set *D* with *N* number of features, there exists *2N* which is the number of candidate subsets. Even with moderate *N*, the search space is exponentially enlarged, and it is prohibited for exhaustive search. Therefore, a lot of strategies have been pro-posed in the literature review, namely sequential search, exponential search, and random search.

(a) Sequential Search *is* an iterative process where only one successor is selected at a time. While this method guarantees completeness, it may not yield an optimal feature subset. To overcome this limitation, variations of the greedy hill-climbing approach have been introduced, including sequential forward selection, sequential backward elimination, and bidirectional selection. Another variation involves alternating between adding and removing features in steps of $k > L$. Although sequential search is straightforward to implement, its search space complexity is *O (N^2)* [76].

(b) Exponential Search is an algorithm used for searching an element in a sorted, unbounded list or array. It is an improvement over the linear search algorithm and combines the features of both binary search and linear search. The basic idea behind exponential search is to start with an initial range that contains the target element, and then keep multiplying the range by a factor (usually 2) until the range covers the entire list. Once the range is determined, a binary search is performed within this range to find the target element.

This approach allows for efficient searching in large arrays and has a time complexity of *O(log n)* in the average case. Exhaustive search guarantees the best solution, but it is not necessarily the most optimized method. Other search algorithms can be used to find optimal solutions with a reduced search space, without sacrificing the quality of the solution. For example, Branch and Bound and Beam Search have been used for smaller subsets, with a search space of *O (N^2).* These algorithms evaluate alternative solutions using heuristic functions, resulting in a more efficient search [36].

(c) Random search is a method that starts with a randomly selected subset and proceeds to obtain an optimal subset. Two strategies can be used to generate the next subset: fully random generation, known as the Las Vegas algorithm, and sequential search, which introduces randomness in the sequential approach. Randomness is introduced to avoid local optima in the search space, which has an order *of O(N2).*

The basic idea behind random search is to randomly select points in the search space and evaluate the function at these points until a satisfactory solution is found. The quality of the solution depends on the size of the search space, the distribution of the random samples, and the number of samples taken. Random search is considered a baseline method for optimization as it provides a simple and straightforward way to search for the optimal solution. The algorithm is particularly useful when the search space is large, the function is complex, and there is no prior knowledge about the function's structure.

Despite its simplicity, random search can sometimes outperform more sophisticated algorithms, especially when the function is highly non-linear and has multiple local optima. The time complexity of random search is *O(N*), where *N* is the number of samples taken [81].

(ii) Successor Generation this is the generation of successor feature subsets that can be achieved using various operators, such as Forward, Backward, Weighing, Random, and Weighting. "For instance, the Forward operator starts with an empty subset X' and iteratively adds a feature $xi \in (X - X')$ that minimizes the validation error $F = argm(X' \cup xi),"$ the small *xi (xi)* represents a specific feature being considered for inclusion in the subset, while the capital *X (X)* represents the set of all available features.

The Forward operator selects a feature xi from the set of available features *X* (excluding those already in the current subset *(X')* that minimizes the validation error. If adding this feature xi to the current subset X' reduces the validation error *(F*) compared to the current error (*E(X')),* then xi is included in the updated subset X'. The operator continues this process until the stopping criteria, such as a predetermined subset size or error threshold, are met.

Backward Feature Selection is a popular method for feature selection in machine learning. It starts with all features included in the set X', and then iteratively removes one feature at a time based on which removal leads to the smallest increase in the error metric F. Specifically, at each step, the algorithm evaluates $(X' - xi)$ for each feature $xi \in X'$ that has not yet been removed. The feature xi that causes the smallest increase in F is removed from X', resulting in a new set $X' = X' - xi$. This process continues until a stopping criterion is met, such as a desired number of features or a threshold value of F.

This method has since been widely used in various applications, including image processing, text classification, and bioinformatics. It is a computationally efficient approach for reducing

the dimensionality of high-dimensional datasets while preserving the most relevant features for accurate modeling and prediction [26].

Compound the main idea of this method is to apply α k number of consecutive forward steps and L number of consecutive backward steps. Based on (X'), forward or backward steps are selected to discover new interactions among features. The stopping criterion should be $K = L$ or $x_i = x_j$. In the sequential feature selection algorithm, it is assured the maximum number of steps with cost $0(n^{K+L+1})$, where $K \neq L$.

Random this strategy comprises all operators. Those that are able to generate a random state in a single step. Other operators are restricted with some criterion such as the number of features or minimizing the error (X') at each step [30].

Weighting this method gives all the features in the solution to the certain degree; where, the search space is continuous. The successor state is a state with different weighting by iteratively sampling the available set of instances [83].

2.9 FEATURE SELECTION METHODS

The following section contains the description of major approaches for feature selection.

2.9.1 Filter Approach

The filter approach to feature selection is a commonly used technique for selecting relevant features based on their intrinsic properties, regardless of the specific ma-chine learning algorithm or model. It involves applying statistical measures or grading methods to individually rank or evaluate each feature, and then selecting the top-ranked features for further analysis or model training.

The general algorithm for a given dataset $D = \{X, L\}$ can be summarized as follows:

Initialize the feature subset X', which can start with options like:

$X' = \{\phi\}$, $X' = \{Null\}$, or $X' \subset X$.

1. Initialize the feature subset X', which can start with options like:

- $X' = \{\phi\}$
- $X' = \{Null\}$

- $X' \subset X$

2. Evaluate each generated subset Xg using an independent measure Im and compare it with the previously determined optimal subset.

3. Iterate the search process until the stopping criterion is met.

4. Output the current optimal feature subset $Xopt$.

Evaluate each generated subset Xg using an independent measure Im and compare it with the previously determined optimal subset.

Iterate the search process until the stopping criterion is met.

Output the current optimal feature subset $Xopt$.

Different filter algorithms can be developed by modifying the subset generator and the evaluation measure [103, 36].

2.9.2 Information Gain Feature Selection

The Information Gain (IG) is a feature selection method that calculates the amount of information obtained for class prediction when the value of a feature is known. The method creates a probabilistic model of a nominal valued feature Y (target class, CHD) by estimating the individual probabilities of the values $y \in Y$ (presence or absence of CHD) from the training data. This model is used to estimate the target class (presence of CHD) for a sample from the training data, and the entropy of the model represents the average number of bits needed to correct the output of the model.

2.9.3 Chi-Square Feature Selection

The Chi-Square statistic is a measure of the deviation from the expected distribution if it is assumed that the occurrence of a feature is independent of the class value. If the Chi-Square score is higher than the critical value determined by the degrees of freedom, then it is concluded that the feature and the class are dependent and such features are selected. The Chi-Square method evaluates features individually by calculating their Chi-Square score in relation to the classes [103, 30].

2.9.4 Embedded Feature Selection

The embedded approach to feature selection is a method that incorporates the feature selection process directly into the model building process. Instead of performing feature selection as a separate step, it is integrated into the algorithm used for training the model.

Here is a general outline of the embedded approach to feature selection:

Choose a machine learning algorithm that inherently performs feature selection or has built-in feature selection mechanisms. Examples include decision trees, random forests, and regularized regression models.

Train the machine learning algorithm using the entire dataset, including all available features. During the training process, the algorithm automatically assigns weights or importance scores to each feature based on their contribution to the model's performance.

Features with low weights or importance scores are considered less relevant and can be discarded. Repeat steps 2-4 using different parameter settings or techniques to fine-tune the feature selection process and improve model performance.

Evaluate the final model on a separate test dataset to assess its performance and generalization ability.

2.10 GENERAL APPROACH TO CLASSIFICATION

Classification is a fundamental task in machine learning, where the objective is to predict the target function that maps a set of features (inputs) to predefined class labels (outputs). The process of data classification comprises two steps: a learning step and a classification step.

In the learning step, a classifier is constructed that describes a pre-determined set of data classes. This is accomplished through a training phase, where a classification algorithm analyzes a well-defined training set consisting of database tuples and their associated class labels. Each tuple is represented as an n-dimensional attribute vector,
$X = (x_1, x_2, x_3, ..., x_n)$, with n measurements made on the tuple from n database attributes, respectively, $A_1, A_2, A_3, ..., A_n$. The class label feature is categorical (or nominal), with each

value serving as a category or class. The training tuples are randomly sampled from the database under analysis, and each tuple is referred to as a sample, example, instance, data point, or object.

The classification model built during the learning step is then used to predict class labels for new data during the classification step. The accuracy of the classification model depends on various factors, including the choice of classification algorithm, feature selection, and preprocessing methods. The success of the classification approach relies on careful consideration and selection of these factors to obtain an accurate and reliable classifier.

The first step in the classification process involves learning a mapping or function, denoted as $y = f(x)$, that can predict the class label y associated with a given tuple X. The goal is to obtain a mapping or function that can effectively separate the data classes, which is typically represented as classification rules, decision trees, or mathematical formulae.

The second step involves using the model for classification, where the predictive accuracy of the classifier is evaluated. If the accuracy is measured using the training set, it may be overly optimistic due to overfitting, where the classifier is sensitive to particular anomalies in the training data that are not present in the general dataset. Therefore, an independent test set is used to evaluate the accuracy of the classifier. The test set is comprised of test tuples with their associated class labels that were not used in constructing the classifier.

The training dataset consists of instances and samples with known class labels, and the classification model is implemented based on the training data. The model is then estimated and tested using the testing dataset, which contains records with unknown class labels. The evaluation of the model's performance is based on the number of testing samples that are correctly predicted, resulting in a confusion matrix.

The accuracy of a classifier on a test set is determined by calculating the percentage of correctly classified test tuples based on the classifier's predicted class. If the accuracy is satisfactory, the classifier can be applied to classify future data tuples with unknown class labels.

To solve a classification problem, the following illustration can be used. Suppose the goal is to classify objects $i = 1,.., n$ into k predefined classes, where k represents the number of classes. For instance, if the aim is to diagnose whether or not a patient is suffering from CHD, then the value of k will be 2, corresponding to the presence or absence of CHD. The available data can be organized as an $n \times p$ matrix X, where xi, j represents the value of feature j in record i. Each row in the matrix X is represented by a vector xi with p features and a class label yi. The classifier can be denoted as (x).

2.11 ENSEMBLE LEARNING TECHNIQUE

Ensemble learning is a machine learning technique that improves the accuracy and stability of predictions by combining the outputs of multiple models. The idea is to train a set of base models on the same dataset, and then aggregate their predictions to obtain a final output. Ensemble methods are often used when a single model may not be sufficient to capture the complexity of the data or to overcome the limitations of a particular algorithm. Ensemble learning can be categorized into homogeneous ensembles, where all base models are of the same type, or heterogeneous ensembles, where base models can be different types. Homogeneous ensembles often use a single base algorithm, such as decision trees or neural networks, while heterogeneous ensembles can use multiple base algorithms. Ensemble learning has been successfully applied to a variety of machine learning problems, including classification, regression, and clustering [24].

The concept and generalization ability of ensemble methods are often stronger than that of base learners, which are also known as weak learners. Ensemble methods have the capability to boost weak learners, which may only be slightly better than random guessing, to strong learners that can make accurate predictions.

The development of ensemble methods can be traced back to three early contributions: combining classifiers, ensembles of weak learners, and mixtures of experts. The combining classifiers approach was primarily studied in the field of pattern recognition, where researchers focused on creating strong classifiers and developing effective combining rules to produce even stronger classifiers. This line of work has resulted in a substantial body of knowledge on the design and implementation of various combining rules [24].

Ensembles of weak learners were mostly studied in the machine learning community. In this approach, researchers focused on weak learners and tried to design powerful algorithms to boost performance from weak to strong [24].

Ensemble formation typically involves two steps: creating the base learners and then combining them. To achieve a good ensemble, it is commonly believed that the base learners should be both accurate and diverse. It is worth noting that the cost of constructing an ensemble is usually not much higher than that of creating a single learner. This is because when using a single learner, multiple versions of it are often generated for model selection or parameter tuning, similar to generating base learners in an ensemble. Additionally, combining base learners is often inexpensive, as most combination strategies are straightforward [24].

Apart from their impressive results in competitions, ensemble methods have been successfully applied to various real-world tasks. They have been found useful in almost all domains where learning techniques are utilized. For example, ensemble methods have greatly benefited computer vision in object detection, recognition, and tracking.

Ensemble methods have shown promise in addressing computer security problems by leveraging information from multiple sources and levels of abstraction. This can be applied ensemble methods to intrusion detection and found that combining the results from ensembles of different feature types improved detection of known attacks. They later proposed an ensemble method for anomaly-based intrusion detection that can detect previously unseen intrusions [24].

2.12 COMMON TYPES OF ENSEMBLE LEARNING

Ensemble learning is a powerful technique in machine learning that combines multiple individual models to make more accurate predictions or decisions. Here are some common types of ensemble learning methods:

(a) ***Bagging (Bootstrap Aggregating):*** In bagging, multiple models are trained on different subsets of the training data, selected through bootstrap sampling with replacement. The final prediction is then determined by aggregating the predictions of all individual models, such as majority voting or averaging.

(b) ***Boosting:*** Boosting works by training models sequentially, where each subsequent model focuses on the examples that were misclassified by the previous models. The predictions of

all models are combined using weighted voting, with more weight given to models that perform better.

(c) ***Random Forest:*** Random forest is an ensemble method that combines bagging with decision trees. It builds multiple decision trees on different subsets of the data and features. The final prediction is made by aggregating the predictions of all decision trees.

(d)Stacking: Stacking involves training multiple models on the same dataset and then using a meta-model to combine their predictions. The meta-model is trained to learn how to best combine the predictions of individual models.

(e) AdaBoost (Adaptive Boosting): AdaBoost is a boosting algorithm that assigns weights to each example in the training data. It trains weak classifiers iteratively, with each subsequent classifier giving more weight to the misclassified examples. The final prediction is determined by combining the predictions of all weak classifiers.

(f) ***Gradient Boosting:*** Gradient boosting is a boosting algorithm that builds an ensemble of models sequentially. Each model is trained to correct the mistakes of the previous model by minimizing a loss function. The final prediction is made by aggregating the predictions of all models.

(h)*Voting:* Voting is a simple ensemble method where multiple models are trained independently, and the final prediction is determined by majority voting (for classification problems) or averaging (for regression problems).

2.13 SELECTED MACHINE LEARNING MODELS

The selected machine learning models in this thesis include:

2.13.1 K-Nearest Neighbor (KNN)

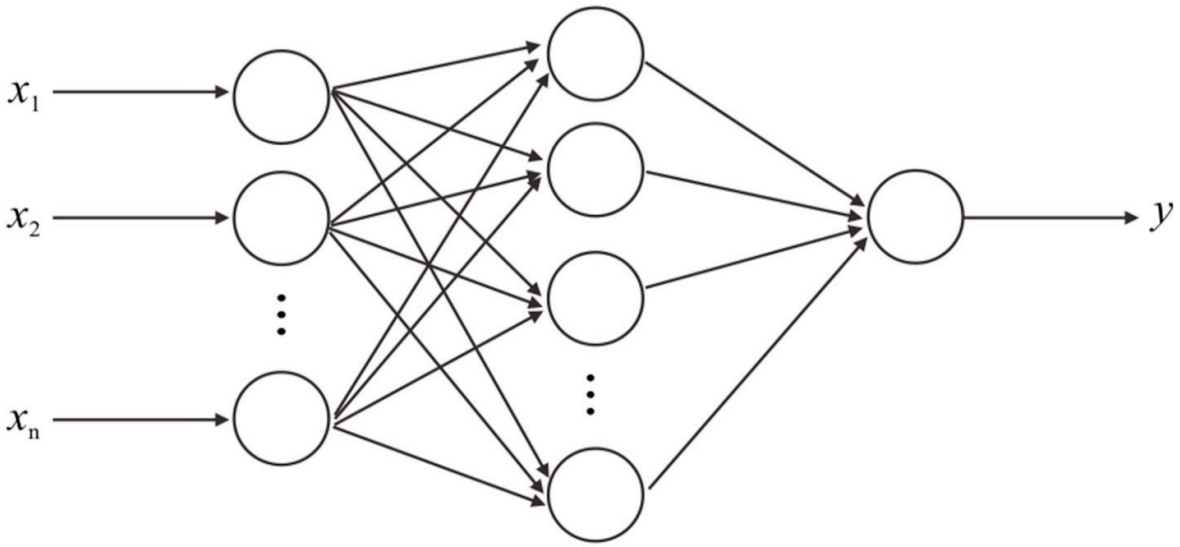
The K-Nearest Neighbor (KNN) algorithm is a machine learning method that is based on memorizing the training set and then using it to predict the label of new instances based on the labels of their closest neighbors in the training set. The idea behind this method is that the features used to describe the domain points are relevant to their labeling in such a way that close-by points are likely to have the same label. In addition, even when the training set is very large, finding the nearest neighbor can be done quickly.

The nearest-neighbor classifiers work by learning by analogy, that is, by comparing a given test instance with similar training instances. The training instances are described by n attributes and each instance is represented by a point in an n-dimensional space.

2.13.2 Multilayer Perceptron (MLP)

Multi-layer Perceptron Neural Networks (MLP-NNs) consist of three sets of layers: the input layer, one or more hidden layers, and the output layer. These networks are composed of neurons, each with biases assigned to them, connections between the neurons, and weights assigned to these connections. The learning process is performed based on the input and target datasets using training algorithms.

The algorithm follows an error-correction rule to adjust the weights and biases of the network, aiming to minimize the error between the target values and the network's output. The input signals, denoted as $x_1, x_2, x_3, \ldots n,$ *are* processed by the network to produce output signals *yk*. The linear combination of the weighted inputs, denoted as *uk*, along with the bias term $bk,$ is passed through an activation function *f* to produce the final output signal. During the training process, the connection weights of each neuron, represented as $w_{k1}, w_{k2}, w_{k3},\ldots wkn,$ are adjusted until the network reaches a minimum allowable error, typically defined by the mean square error (MSE) function.



Input Layer Hidden Layer Output

Figure 5 Architecture of a Multilayer Perceptron Neural Network [100]

2.13.3 Support Vector Machine (SVM)

SVM is a statistical method developed by Guyon and Vapnik that is widely used for various classification tasks [45]. The algorithm works by transforming the original data from a low-dimensional space to a high-dimensional space using nonlinear mapping, and then finding the optimal linear separating hyperplane with the maximum margin. SVM is based on the principles of classical statistical learning theory and has been found to have good generalization performance with an interpretable model.

From a statistical learning perspective, the motivation for using SVM as a binary classifier comes from the bounds on the generalization error. These bounds have two important properties: the upper bound is independent of the size of the input space, and the bound is minimized by maximizing the margin between the hyperplane separating the two classes and the closest data points to each class, known as support vectors. Support vectors are called so because they support the position of the hyperplane [25].

2.14 RELATED WORKS TO DIAGNOSE OF CORONARY HEART DISEASE

Related researches have investigated several methods and procedures to increase precision and dependability in the detection of coronary heart disease. Some noteworthy research includes:

(a) Techniques and methodologies for creating a successful medical decision support sys-tem. They introduced a method that utilized the Statistical Analysis System (SAS) base software for diagnosing coronary heart disease and employed an ensemble model of neural networks. By combining multiple individual neural networks trained on the same task, the method aimed to improve generalization performance. However, the resulting system was not highly accurate [9].

(b) Hybrid approach that combines Genetic Algorithms (GAs) and Support Vector Machines (SVMs), referred to as the wrapper approach. The GA component selects the best attributes in the data set, while SVM classifies the patterns based on the reduced data set. The experiments were conducted using datasets from the UCI machine learning repository. The results showed that the GA-SVM hybrid was effective in classification after

removing irrelevant attributes, with an accuracy of 76.20%. The approach was also applied to multi-class domains, resulting in an average accuracy of 84.07%. However, this accuracy remains low [19].

(c) Generalized Regression Neural Network (GRNN) and a Radial Basis Function Network (RBFN) to diagnose coronary heart disease. The study used data from 200 patients for training and testing the classifiers, with the SVM classification method utilized as the GRNN. The results showed that the RBF Network produced more accurate medicine prescriptions, verified by a doctor, compared to the SVM which provided unsatisfactory results. The researchers concluded that with more training data, a better performance could be expected [34].

(d) Classification method that involved preprocessing the data with Principal Component Analysis (PCA) prior to using the classification model for heart disease diagnosis. They utilized classical Electronic Medical Record (EMR) heart data sets. The authors emphasized the importance of using an effective global optimization technique, such as differential evolution, in fitting the classification model instead of local optimization approaches. Their findings indicated that preprocessing the data with PCA improved the classification accuracy. Despite achieving an average accuracy of 82%, the proposed system was still considered low [36].

(e) By suggesting the Extreme Learning Machine (ELM) classifier for categorizing ECG patterns, machine learning is proposed for the diagnosis of cardiac illness. Using information from the Physio Net Arrhythmia database, the researchers tested their ELM classifier, comparing the outcomes to those of support vector machine classification. The outcomes showed that the ELM classifier performed better than the support vector machine classifier in categorizing five different categories of normal and abnormal ECG frequencies. The k-nearest neighbor classifier (kNN) and the radial basis function neural network classifier (RBF) were built into the ELM classifier. However, implementation on any platform was not included in the study [59].

(f) Web-based and Fuzzy Logic Expert System for heart failure disease diagnosis. The system consisted of a Knowledge Base (database), a Fuzzy Logic component, a Fuzzy

Inference Engine, and a Decision Support Engine that included a cognitive and emotional filter, as well as telemedicine facilities. The system was implemented using Hypertext Preprocessor (PHP), JavaScript, and Hypertext Markup Language (HTML) with My Structured Query Language (MySQL) as the database management system. The performance of the system was analyzed using data from selected heart failure patients and a survey of heart failure disease experts at the State Specialist Hospital in Akure, Nigeria. The results showed satisfactory performance, but the authors noted that the system was based on a small number of clinical instances and that ensemble learning could have improved its performance [2].

(g) Diagnosing Coronary Heart Disease using Ensemble Machine Learning. The researchers developed an advanced ensemble machine learning technology using an adaptive Boosting algorithm for accurate coronary heart disease diagnosis and outcome predictions. The ensemble learning classification and prediction models were applied to four different datasets, including patients diagnosed with heart disease from the Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology (HIC), Long Beach Medical Center (LBMC), and Switzerland University Hospital (SUH). The results showed that the developed models achieved an average accuracy of 80.14%, surpassing the accuracy of previous research. However, the authors noted that feature selection could have been used to extract the most relevant features, which would have improved the model's accuracy [48].

(h) A comprehensive, well cited coverage of the field makes this book a valuable reference for any researcher. The book provides a mathematical description of a comprehensive set of deep learning algorithms, but could benefit from more pseudocode examples. The authors provide an adequate explanation for the many mathematical formulas that are used to communicate the ideas expressed in this book. The lack of both exercises and examples in any of the major machine learning software packages makes this book difficult as a primary undergraduate textbook [93].

3 SYSTEM STRUCTURE OF PREDICTIVE MODELS FOR CORONARY HEART DISEASE

The creation of a collection of prediction models for CHD diagnosis and the system architecture, the methods utilized, stages of model creation are all described in details

3.1 System Architecture

The system architecture for CHD diagnosis includes key steps such as data collection and preprocessing, model development, and model validation. The historical dataset containing diagnostic indicators was subjected to filter-based feature selection to identify relevant features. The reduced feature set was then used for training and testing supervised machine learning algorithms. The performance of individual classifiers and ensemble learning models was evaluated for predicting CHD presence or absence. Figure 7 depicts.

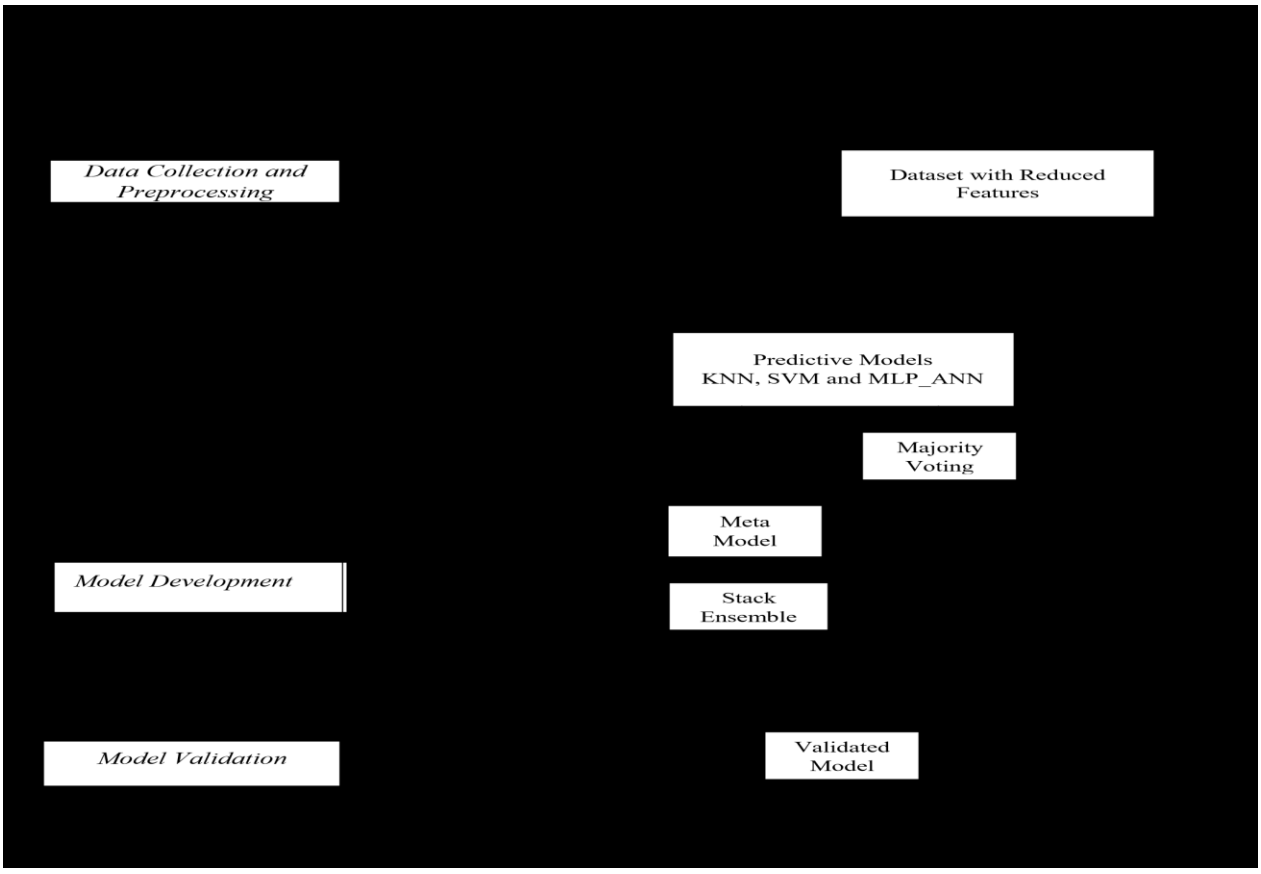


Figure 6 The System Architecture

3.2 DATA IDENTIFICATION AND COLLECTION

It involves identifying the relevant data sources and collecting the necessary data to address a specific research question or objective. The following steps outline the process of data identification and collection:

3.2.1 Data Identification

After conducting a comprehensive review of the existing literature on CHD, several key features were identified to be evaluated in patients. The monitored variables, as noted in prior studies, were compared with the variables monitored by the cardiologists at University College Hospital (UCH).

3.2.2 Data Collection

Disease Databases available in the UCI Machine Learning Repository [6].These databases contain data on CHD clinical instances, contributed by the Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology (HIC), Long Beach Medical Center (LBMC), and University Hospital in Switzerland (SUH), respectively. The databases contain 303, 294, 200 and 123 clinical instances in each data set, respectively. This results in a total combination of 920 clinical instances.

Each heart disease database has the same clinical instance format for each patient. Each clinical instance contains a total of 14 attributes and one target attribute. The target attribute refers to the status of the presence of coronary heart disease in the patients. It is represented by an integer value “0” or “1,” where value “0” signifies absence and the value “1” signify the presence of CHD. Experimental data can be found in Table 1 and Table 2 respectively.

[6] (https://archive.ics.uci.edu).

Table 1 Sample of the CHD Dataset Used [70]

age	sex	Cp	trest bps	Chol	Fbs	restec g	thal	exang	Old peak	Slope	ca	thal	Class
63	male	Typical angina	145	23333	true	LVH	150	No	2.3	Down sloping	0	fixed defect	No
67	male	Asymptomatic	160	286	false	LVH	108	Yes	1.5	Flat	3	Normal	Yes
67	male	Asymptomatic	120	229	false	LVH	129	Yes	2.6	Flat	2	reversible defect	Yes
37	male	Non-angina	130	250	false	normal	187	No	3.5	Down sloping	0	Normal	No
41	female	Atypical angina	130	204	false	LVH	172	No	1.4	Up sloping	0	Normal	No
56	male	Atypical angina	120	236	false	normal	178	No	0.8	Up sloping	0	Normal	No
62	female	Asymptomatic	140	268	false	LVH	160	No	3.6	Down sloping	2	Normal	Yes
57	female	Asymptomatic	120	354	false	normal	163	yes	0.6	Up sloping	0	Normal	No
63	male	Asymptomatic	130	254	false	LVH	147	no	1.4	Flat	1	reversible defect	Yes
53	male	Asymptomatic	140	203	true	LVH	155	yes	3.1	Down sloping	0	reversible defect	Yes
57	male	Asymptomatic	140	192	false	normal	148	no	0.4	Flat	0	fixed defect	No
56	female	Atypical angina	140	294	false	LVH	153	no	1.3	Flat	0	Normal	No

Table 2 Description of the Variables Identified for CHD [70]

Variable Name	Labels
Age of Patient (in years) – age	Integer
Sex of Patient – sex	Male Female
Chest Pain Type – cp	Typical Angina Atypical Angina Non-angina Asymptomatic
Resting Blood Pressure (in mmHg) – trestbps	Integer
Serum Cholesterol (in mg/dl) – chol	Integer
Fasting Blood Sugar (> 120 mg/dl) – fbs	True False
Resting ECG Results – restecg	Normal ST-T Wave Abnormality Probable Left Ventricular Hypertrophy
Maximum Heart Rate achieved (in bpm) – thalach	Integer
Exercise Induced Angina – exang	Yes No
ST depression induced by exercise relative to rest – oldpeak	Real
Slope of Peak ST Segment – slope	Up sloping Flat Down sloping
Number of Major Vessels colored by fluoroscopy – ca	Integer
Heart Rate – thal	Normal Fixed Defect reversible Defect
Presence of Coronary Heart Disease (CHD)	No Yes

3.3 DATA PRE-PROCESSING

The preprocessing of data is an essential step in the process of gaining knowledge. Real-world data often contains missing information, noise, and inconsistencies. In this study, the data preprocessing tasks included cleaning, discretization, and normalization.

The data cleaning process aimed to fill in missing values, remove noise, identify outliers, and correct inconsistencies. In this study, records with missing data were discarded and default values were used where necessary. After the cleaning process, 118 instances were removed due to a high number of missing attributes, and 802 instances were found to be suitable for the study.

Data discretization is considered a crucial data preprocessing task. It transforms nominal data into numeric data, reducing noise, smoothing data, reducing data size, and enabling specific methods that use nominal data. In this thesis, nominal input values (attributes) were converted into numeric values through data discretization.

The normalization of data is crucial for accurate classification. Normalizing the input values for each attribute in the training dataset helps to speed up the learning process. To prevent one attribute from overpowering the others in terms of distance measurement, all the input features in the dataset were normalized before the training and testing process. Min-Max normalization was used in this thesis, which adjusts the values to a range between 0 and 1. Equation (1) presents the Min-Max normalization formula applied.

$$x\_normalized = (x - min(x)) / (max(x) - min(x))$$

Where,

x = new value for the variable x

x_{min} = current value for variable x

x_{min} = minimum value in the dataset

x_{max} = maximum value in the dataset

3.4 EVOLUTION OF ENSEMBLE LEARNING MODELS

An effective method for combining various separate models to get more reliable and accurate predictions is ensemble learning models.

KNN, SVM, MLP (stacked), and voting as the ensemble learning models in this thesis is to leverage the strengths and diversity of these models to improve the prediction accuracy.

3.4.1 The Concept of Majority (Hard) Voting

The Voting ensemble method involves aggregating the decisions made by multiple classifiers. The process starts by dividing the training data into smaller equal parts and building a classifier for each subset of data. The simplest form of voting is based on majority or plurality voting, where each classifier casts one vote. The final prediction is based on the class that receives the most votes, meaning the class with the highest aggregate of votes is chosen as the final decision and the system architecture can be seen in the Figure 8 below.

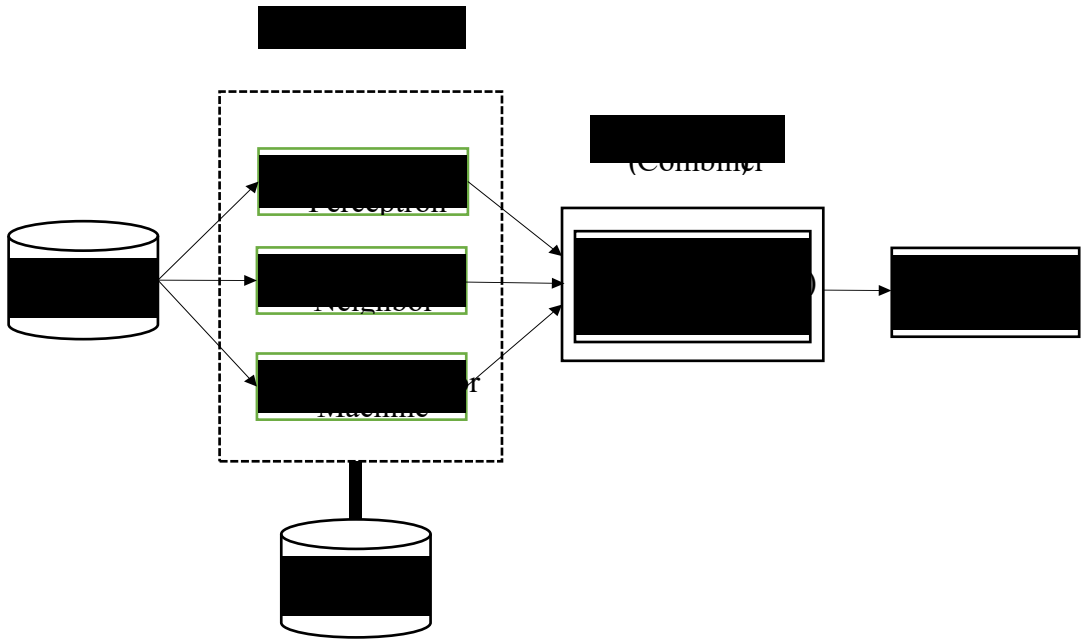


Figure 7 Architecture of Majority Voting Ensemble [103]

3.4.2 The Concept of Stacked Generalization (Stacking)

A method used in ensemble learning called stacked generalization, commonly referred to as stacking, involves training many models and combining their predictions using a meta-model. Stacking is a technique used to combine the advantages of various models to provide predictions that are more reliable and accurate and details can be found in Figure 9.

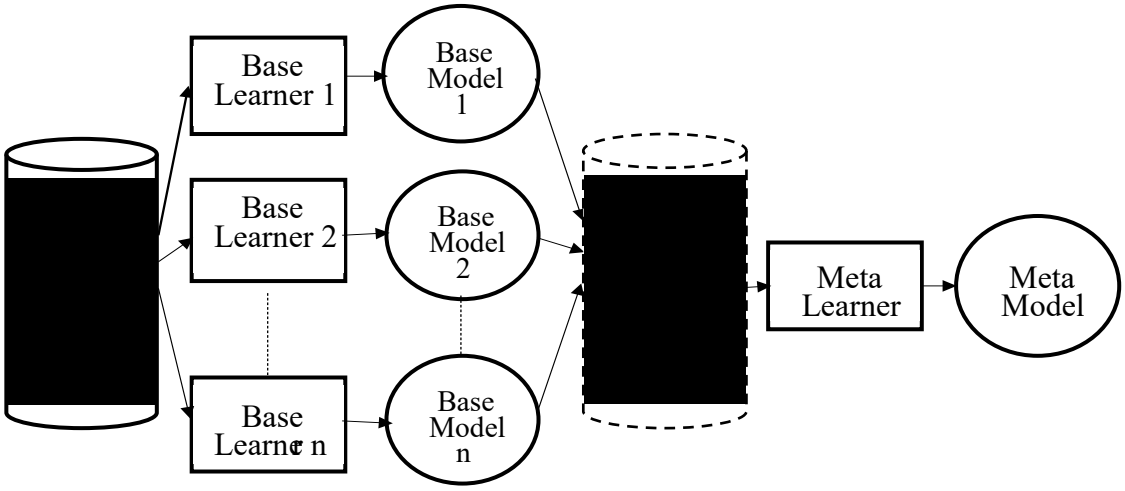


Figure 8 Architecture of Stacked Generalization Ensemble

3.4.3 Stacked Generalization Approach for CHD Diagnosis

Three machine learning techniques or methods are used, and it comprises of K-Nearest Neighbor (KNN), Multilayer Perceptron (MLP) and Support Vector Machine (SVM) as the base learners to a combiner (Meta) learner which is logistic regression using stacked generalization algorithm.

3.5 PERFORMANCE MEASURES

To reduce bias in the random sampling of training and test data, k-Fold Cross Validation was employed. This involves randomly dividing the initial data into k mutually exclusive subsets or "folds" - $D_1, D_2,..., D_k$ - of approximately equal size. Training and testing are repeated k times. Extensive tests across numerous datasets and learning techniques suggest that 10 folds provide the most accurate estimate of error, with some theoretical support. Despite ongoing debates about the best evaluation scheme in the field of machine learning,

10-fold cross-validation has become the standard approach in practical applications. Furthermore, using stratification has been shown to marginally enhance results. Consequently, stratified 10-fold cross-validation is the standard evaluation technique in scenarios where limited data is available.

The following metrics are generally included in the performance metrics of ensemble models, notably for the KNN, SVM, and MLP classifiers:

Accuracy: The ratio of correctly categorized instances to all instances in the dataset is measured. Better performance is indicated by a higher accuracy.

Precision: It shows how well the model can pick out positive examples from the entire set of projected positive examples. The ratio of true positives to the total of true positives and false positives is used to compute it.

Recall (Sensitivity): It gauges how well the model is able to distinguish between the total numbers of true positive cases. The ratio of true positives to the total of true positives and false negatives is used to compute it.

3.6 10-FOLD CROSS VALIDATION

By dividing the data into k subsets or folds, a machine learning model's performance is assessed using the K-fold cross-validation technique. Each fold serves as the testing set once, while the remaining k-1 folds are utilized as the training set, during the model's k times of training and evaluation. Repeating the procedure gives each fold an opportunity to serve as the testing set.

Following are the steps involved in K-fold cross-validation:

To divide the dataset: The original dataset is split into k folds of equal size.

The remaining k-1 folds are utilized as the training set, and one fold is chosen as the testing set for each iteration. The training set is used to develop the model, and the testing set is used to assess it.

Performance measures: The accuracy, precision, recall, F1 score, and other assessment metrics are used to assess the model's performance.

Combining the outcomes: To achieve a more reliable estimate of the model's performance after all iterations, the performance measures from each fold are averaged.

Compared to a single train-test split, K-fold cross-validation makes it possible to evaluate the model's performance in a more accurate and objective way. It offers a more accurate prediction of the model's performance on unknown data and aids in identifying problems like under fitting or overfitting. K-fold cross-validation offers a more thorough assessment of the model's generalization skills by repeating the training and testing procedure with several subsets of data.

II. ANALYSIS

4 RESULT

This section covers the experimental setup, result obtained as well as discussion of the results obtained from the individual and ensemble models.

4.1 EXPERIMENTAL SETUP

For the ease of machine learning and the fact that it is much better, easier to work with numbers in programming, the discussion attribute value Present or Absent were consequently converted to integers 1 and 0 respectively. After both the training and testing sets were formatted into acceptable format, classification experiments were then carried out. In order to achieve the stated objectives of the thesis, the experimental setup was broken into two major steps and the results obtained are illustrated in form of Tables and Charts.

The steps are predefined as follows.

In Step 1 : Each of the two feature selection algorithms (Information Gain and Chi Squared) were used independently to identify main features among the initially identified variables in the data set collected from University of California, Irvine (UCI) machine learning repository regarding the diagnosis of CHD. The two-feature selection algorithm were individually tested on the testing data set.

In Step 2, predictive models were created using 10-fold cross-validation and the three fundamental algorithms of Nearest Neighbor (KNN), Support Vector Machine (SVM), and Multilayer Perceptron (MLP), as well as ensemble learners as shown in the Figure 10.

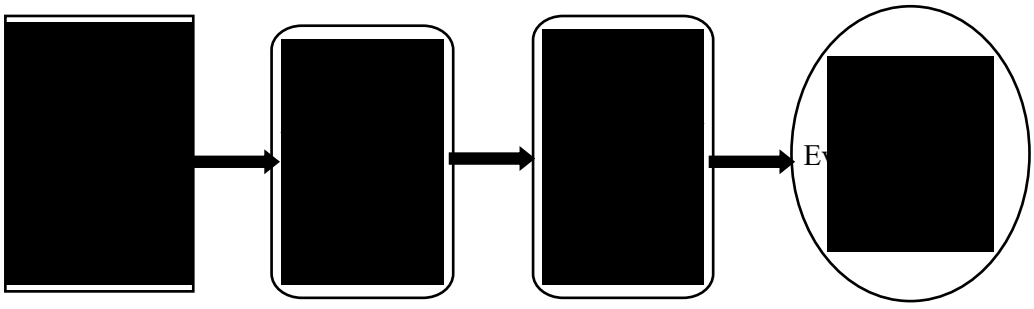


Figure 9 Chain of Operation

In Step 3: The stacked generalization ensemble learning algorithm to combine three base (Level-0) classifiers obtained from Step 1. Logistic Regression as a Meta (Leve-1) classifier for each of the base classifiers. The resulting stacking ensemble model was tested on both the training set and the testing set.

In Step 4: The results obtained from Step 1 to Step 3 and selected the best Ensemble learning classification model that provided the highest performance accuracy. It is important to note that the three steps were independent of one another in the experimental setup.

Step 5: CHD prediction results of the stacked generalization approach with another ensemble learning method, the Hard Voting Ensemble is compared.

4.2 RESULTS OF FEATURE SELECTION METHODS.

After identifying and describing the CHD diagnosis dataset, the next crucial step was to identify the most relevant features that would improve the accuracy of predicting CHD diagnosis. In order to achieve this, two filter-based feature selection methods were employed. The ranked features were then grouped into three categories based on their rank: the first group contained the top 5 features, the second contained the top 10, and the last group contained all 13 features. These grouped features were then passed to each classifier for classification, and the group that performed best was chosen as the most relevant features. The group of 10 features that performed the best was selected and used for the design of the models.

4.2.1 Result of Information Gain Feature Selection Method

The information gain feature selection algorithm was utilized in this thesis to identify the most relevant features by computing the entropy of each attribute and the diagnosis of CHD, and then measuring their difference to determine the information gain of each attribute. A higher information gain value indicates a more important attribute. The performance of the models with the information gain feature selection technique demonstrated that the best accuracy was achieved when the top 10 ranked attributes were chosen. This approach evaluates the amount of information in bits related to the class prediction, assuming that the only available information is the presence of a feature and its corresponding class distribution. The ranked features based on their information gain values are presented in the

Table3 and Figure 11. The feature selection process involved ordering the features in ascending order based on their magnitude or values.

Table 3 Ranked Features Based on the information Gain Values

S/N	**Features**	**Information Gain Values**
1	Chest Pain (CP)	0.14904957856047152
2	Number of major vessels colored by fluoroscopy (ca)	0.1476289555070711
3	Maximum heart rate achieved (thalach)	0.11526320858026606
4	ST depression induced by exercise relative to rest (oldpeak)	0.10432160094068466
5	Sex of patient	0.09976977634846595
6	Exercise induced angina (exang)	0.05428766927091733
7	Age of patient	0.04668300137776771
8	Heart rate (thal)	0.026188531278747762
9	Slope of Peak ST segment (slope)	0.01707976730558536
10	Resting blood pressure (trestbps)	0.016608304886700287
11	Serum Cholesterol (chol)	0.015764238505083572
12	Fasting blood sugar (fbs)	0.01386194518003192
13	Resting ECG result (restecg)	0.0

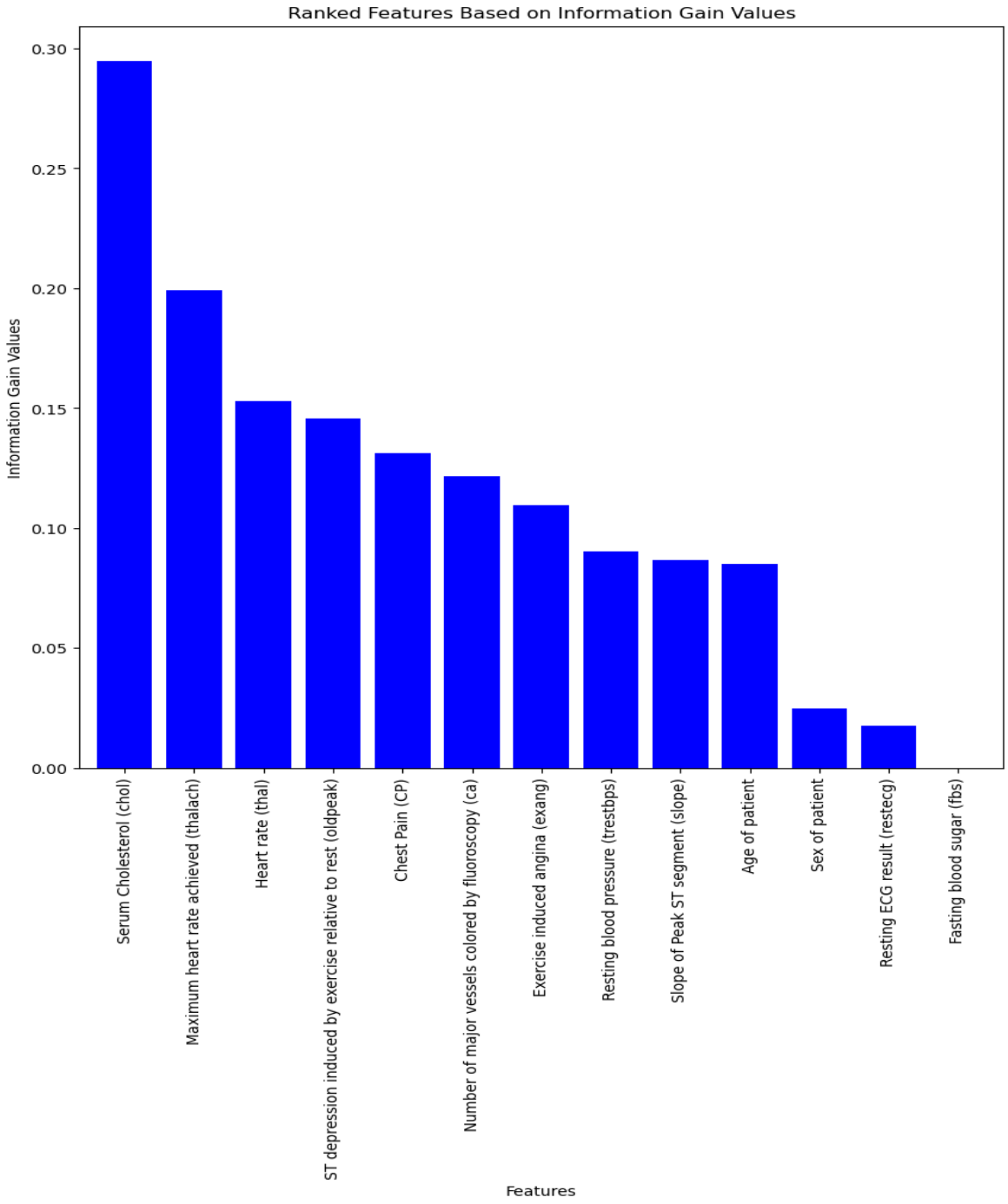


Figure 10 Ranked Features Based on the Information Gain Values

4.2.2 Result of Chi-Square Feature Selection Method

To select features for the dataset, the Chi-Squared feature selection algorithm was employed. This algorithm evaluates whether the occurrence of a specific term and a specific class are independent. Table 4 and Figure 12 display the ranked features based on their p-values. The feature selection process involved sorting the features in increasing order of their magnitude or values, and then selecting the top features.

Table 4 Ranked Features Based on Chi-Square Test Score.

S/N	Features	Chi-Square Test Score
1	Exercise induced angina (exang)	25.17454350161117
2	Heart rate (thal)	22.920031255745542
3	Number of major vessels colored by fluoroscopy (ca)	20.233812949640267
4	Chest pain (cp)	14.356000025841295
5	Sex of patient (sex)	6.331969036629627
6	ST depression induced by exercise relative to rest (oldpeak)	3.451188338458821
7	Slope of Peak ST segment (slope)	2.385997400909677
8	Maximum heart rate achieved (thalach)	1.8678003094462325
9	Age of patient (age)	0.7885741958306334
10	Resting ECG result (restecg)	0.49438380003164023
11	Resting blood pressure (trestbps)	0.34372703460420395
12	Serum Cholesterol (chol)	0.0036980195630211756
13	Fasting blood sugar (fbs)	0.0036014405762304527

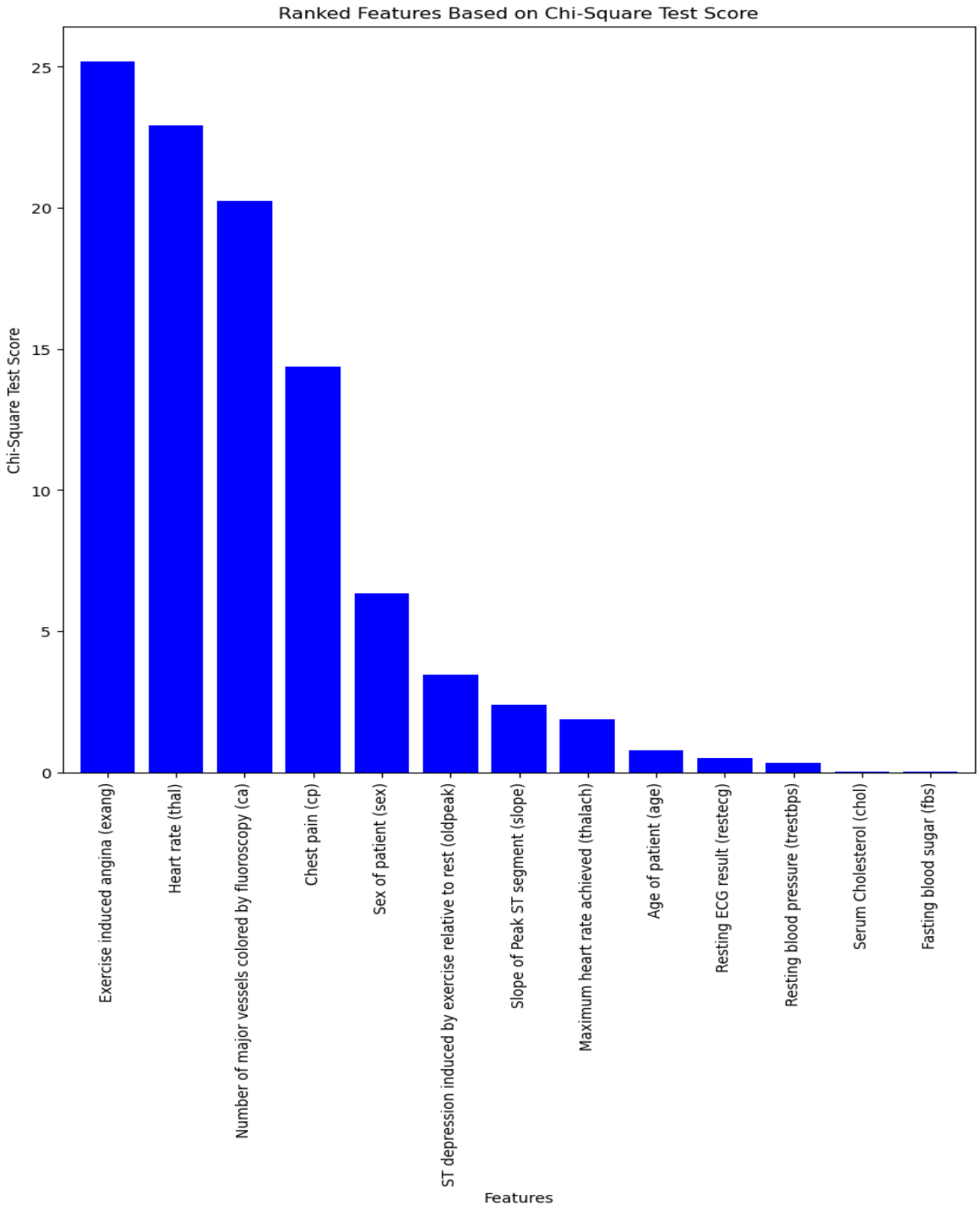


Figure 11 Ranked featurees Based on Chi square Test score

4.3 RESULTS OF CLASSIFICATION

Ensemble machine learning methods were employed in this study to forecast coronary heart disease. Individual prediction models were created to accomplish this utilizing K-Nearest Neighbor, Multilayer Perceptron Neural Network, and Support Vector Machine, three supervised machine learning methods. Both a stacked generation ensemble model

and a voting ensemble strategy were used to integrate these models. The Meta model was created using logistic regression and similar experiment is done on all the three studied classifier.

Five experiments were done to gauge how well the models worked. Each experiment took into account three different circumstances and used different sets of attributes chosen using information gain and chi-square feature selection techniques. Based on the many trials and circumstances, 15 distinct models were created in all.

A training dataset with 802 instances was used for the trials, and 10-Fold Cross Validation was employed to randomly sample the data for training and testing. Standard measures including accuracy, precision, recall, and F-measure, which were calculated using the Confusion Matrix, also known as the predictive classification table, were used to evaluate the models' performance.

4.3.1 Model Building Using K-Nearest Neighbor (KNN)

The predictive classification is analyzed in Table 5 and Table 6 shows detailed performance measures:

Table 5 Confusion Matrixes for Experiment 1 (KNN Model)

		Predicted	
Model	Actual	Yes	No
KNN with all attributes	Yes	318(39.7%)	78(9.7%)
KNN with all attributes	No	79(39.7%)	327(40.8%)
KNN with Information Gain Feature Selection	Yes	332(41.40%)	64 (8%)
KNN with Information Gain Feature Selection	No	62(7.7%)	344(42.9%)
KNN with Information Chi Square Selection	Yes	335(41.8%)	61(7.6%)
KNN with Information Chi Square Selection	No	57(7.1%)	349(43.5%)

Table 6 Detailed Performance Measures for Experiment 1 (KNN Model)

Model	**KNN with all Attributes**	**KNN with Information Gain FS**	**KNN with Chi-Square FS**
Correct Classification	645 (80.4%)	676 (84.3%)	684 (85.3%)
Accuracy (%)	80.4	84.3	85.3
Precision (%)	80.1	84.3	85.5
Sensitivity (%)	80.3	83.8	84.6
Specificity (%)	80.7	84.3	85.1
FAR (%)	19.3	15.7	14.9
F-1 Score (%)	80.2	84.1	85

The K-Nearest Neighbor (KNN) models were evaluated on different attribute sets and feature selection methods. In the first KNN model using all 13 attributes, the overall accuracy rate was moderately low at 80.4%. The model performed better in identifying patients with coronary heart disease (sensitivity of 80.3%) than identifying patients without the disease (sensitivity of 80.7%). The precision score and F-Measure value indicated a balanced trade-off between precision and recall.

In the second KNN model using 10 attributes selected by the information gain method, the overall accuracy rate improved to 84.3%. The model showed improved sensitivity for both positive and negative cases, with a precision score and F-Measure value indicating a balanced performance.

The third KNN model using 10 attributes selected by the chi-square method achieved the highest accuracy rate of 85.3%. It showed a similar pattern of improved sensitivity for positive and negative cases, with a higher precision score and F-Measure value compared to the previous models.

Overall, the KNN models performed better when using the selected attributes, with the chi-square feature selection method resulting in the highest accuracy rate. The classification accuracy increased from 80.4% (with all attributes) to 84.3% (with information gain) and further to 85.3% (with chi-square).

4.3.2 Model Building Using Multilayer Perceptron

Table 9 and Table 10 illustrate the performance measures obtained from the predictive classification experiment.

Table 7 Confusion Matrixes for Experiment 2 (MLP Model)

		Predicted	
Model	Actual	Yes	No
MLP with all attributes	Yes	326(40.7%)	70(8.7%)
MLP with all attributes	No	64(8%)	342(42.6%)
MLP with Information Gain Feature Selection	Yes	351(43.8%)	45(5.6%)
MLP with Information Gain Feature Selection	No	42(5.2%)	364(45.4%)
MLP with Chi Square Selection	Yes	356(44.4%)	40(5%)
MLP with Chi Square Selection	No	39(4.9%)	367(45.8%)

Table 8 Detailed Performance Measures for Experiment 2 (MLP Model)

Model	**MLP with all Attributes**	**MLP with Information Gain FS**	**MLP with Chi-Square FS**
Correct Classification	668 (83.3%)	715 (89.2%)	723 (90.2%)
Accuracy (%)	83.3	89.2	90.1
Precision (%)	83.6	89.3	90.1
Sensitivity (%)	82.3	88.6	89.9
Specificity (%)	83	89	90.2
FAR (%)	17	11	9.8
F-1 Score (%)	83	89	97

The first MLP model, trained on all 13 attributes, achieved an accuracy of 83.3%. It showed a sensitivity of 82.3% in correctly identifying patients with coronary heart disease and a sensitivity of 83% in correctly identifying patients without the disease. The model had an average precision of 83% and a balanced F-measure of 83%.

The second MLP model, built with 10 attributes selected through information gain, achieved an accuracy of 89.2%. It exhibited a sensitivity of 88.6% for patients with the disease and a specificity of 89% for patients without the disease. The model had an average precision of 89.15% and a balanced F-measure of 89%.

The third MLP model, built with 10 attributes selected through chi-square, achieved the highest accuracy of 90.1%. It showed a sensitivity of 89.9% for patients with the disease and a specificity of 90.2% for patients without the disease. The model had an average precision of 90.15% and a balanced F-measure of 90%.

Overall, the MLP models performed better when using the selected attributes, with the model built with chi-square feature selection achieving the highest accuracy of 90.1%.

4.3.3 Model Building Using Support Vector Machine (SVM)

Table 9 and Table 10 illustrate the performance measures obtained from the predictive classification experiment.

Table 9 Confusion Matrixes for Experiment 3 (SVM Model)

		Predicted	
Model	Actual	Yes	No
SVM with all attributes	Yes	318(39.7%)	78(9.7%)
SVM with all attributes	No	72(9%)	334(41.7%)
SVM with all Information Gain Feature Selection	Yes	344(42.8%)	52(6.5%)
SVM with information Gain Feature Selection	No	50(6.2%)	356(44.4%)
SVM with Chi Square Selection	Yes	347(43.4%)	49(6.1%)
SVM with Chi Square Selection	No	39(4.9%)	360(44.8%)

Table 10 Detailed Performance Measures for Experiment 3 (SVM Model)

Model	**SVM with all Attributes**	**SVM with Information Gain FS**	**SVM with Chi-Square FS**
Correct Classification	653 (81.4%)	700 (87.3%)	707 (88.2%)
Accuracy (%)	81.3	87.3	88.2
Precision (%)	81.5	87.3	88.3
Sensitivity (%)	80.3	86.9	87.6
Specificity (%)	81.1	87.3	88
FAR (%)	18.9	12.7	12
F-1 Score (%)	80.9	87.1	88

The SVM Classifier was evaluated in three scenarios: using all 13 attributes, using 10 attributes selected by information gain, and using 10 attributes selected by chi-square feature selection. The first SVM model achieved an accuracy rate of 81.3% with all attributes, while the second and third models achieved accuracy rates of 87.3% and 88.2% respectively with the selected attributes. The models showed moderately low overall accuracy rates but demonstrated a balanced precision and recall. The SVM model using chi-square feature selection had the highest accuracy among the three scenarios and chi square had 88.2%.

4.3.4 Model Building Using Voting Ensemble (Hard Voting)

The ensemble result based on performance can be found below:

Table 11 Detailed Performance Measures for Experiment 4 (Voting Ensemble Model)

Model		**Voting Ensemble with all Attributes**	**Voting Ensemble with Information Gain FS**	**Voting Ensemble with Chi-Square FS**
Correct Classification		722 (90%)	738 (92%)	746 (93%)
Accuracy (%)		90	92	93
Precision (%)		90.3	92	93.4
Sensitivity (%)		89.4	91.9	92.4
Specificity (%)		89.8	92.1	92.7
FAR (%)		10.2	7.9	7.3
F-1 Score (%)		91.9	94.9	97

4.3.5 Performance Measures of Models Using Stacked Ensemble

To evaluate the performance of the stacked ensemble classifier in predicting coronary heart disease. Three scenarios were considered: using all 13 attributes, using 10 attributes selected through information gain feature selection, and using 10 attributes selected through chi-square feature selection. The objective was to analyze the impact of attribute selection on the models' performance.

In scenario one, the stacked ensemble classifier was trained on the complete dataset of 802 instances with all 13 attributes. In scenario two, the classifier was trained on the complete dataset with only 10 selected attributes using the information gain feature selection method. In scenario three, the classifier was trained on the complete dataset with only 10 selected attributes using the chi-square feature selection method.

Performance of the stacked ensemble using KNN, SVM and MLP:

Table 12 Detailed Performance Measures for stacked ensembles

Model	**Stacked Ensemble with all Attributes**	**Stacked Ensemble with Information Gain FS**	**Stacked Ensemble with Chi-Square FS**
Correct Classification	738 (92%)	762 (95%)	778 (97%)
Accuracy (%)	92.02	95.01	97.00
Precision (%)	91.9	95.4	97.2
Sensitivity (%)	91.9	94.4	96.07
Specificity (%)	92.1	94.6	96.8
FAR (%)	7.9	5.4	3.2
F-1 Score (%)	91.9	94.9	97

The first stacked ensemble model, built with all 13 attributes, achieved an overall accuracy of 92%. It correctly classified 738 instances and misclassified 64 instances. The model had a sensitivity of 91.9% in identifying patients with coronary heart disease and a specificity of 92.1% in identifying patients without the disease. The precision score was 92% and the False Alarm Rate was 7.9%. The F-measure value indicated a balanced precision and recall.

The second voting ensemble model, using 10 attributes selected through information gain, achieved an overall accuracy of 95%. It correctly classified 754 instances and misclassified

40 instances. The model had a sensitivity of 94.4% and a specificity of 94.6%. The precision score was 96% and the False Alarm Rate was 6.54%. The F-measure value indicated a balanced precision and recall.

The third voting ensemble model, using 10 attributes selected through chi-square, achieved an overall accuracy of 97%. It correctly classified 778 instances and misclassified 24 instances. The model had a sensitivity of 96.7% and a specificity of 96.8%. The precision score was 97% and the False Alarm Rate was 3.2%. The F-measure value indicated a balanced precision and recall.

Overall, the stacked ensemble model performed better on the selected attributes, with the model using 10 attributes selected through chi-square achieving the highest accuracy of 97%. The accuracy increased from 92% (using all attributes) to 95% (using information gain) and then to 97% (using chi-square).

4.4 ANALYSIS OF RESULTS

Medical examination interpretations are dependent on the observer (i.e., medical experts). Therefore, a predictive model was developed to diagnose coronary heart disease with less dependence on the observer, particularly when a cardiologist is not readily available.

To determine optimal algorithms for predicting coronary heart disease, five experiments were conducted and their classification performance was compared. The experiments had two main objectives: to observe how attribute selection affects classification accuracy and to compare the performance of KNN, MLP, SVM, Voting ensemble, and stacked ensemble classifiers.

4.4.1 Effect of Attribute Selection

In all experiments, three scenarios were evaluated: one containing all 13 attributes, another containing 10 attributes selected using information gain feature selection method, and the third containing 10 selected attributes using chi-square feature selection method. The results indicated that reducing the number of attributes improved classification accuracy.

Attribute selection played a significant role in improving the models by increasing classification accuracy and reducing model complexity by eliminating irrelevant attributes. Addi-

tionally, having a reduced number of attributes resulted in faster execution time. The algorithms showed improved classification accuracy when using selected attributes, as the removed attributes were not relevant in predicting coronary heart disease.

Table 13 The Effect of Attribute Selection on Classification Accuracy (%)

Model	**All the attributes**	**10 attributes with Info_Gain FS**	**10 attributes with Chi-Square FS**
KNN	80.42	84.29	85.29
MLP	83.30	89.15	90.15
SVM	81.42	87.28	88.15
Voting Ensemble	90.02	92.02	93.02
Stacked Ensemble	92.02	95.01	97.00

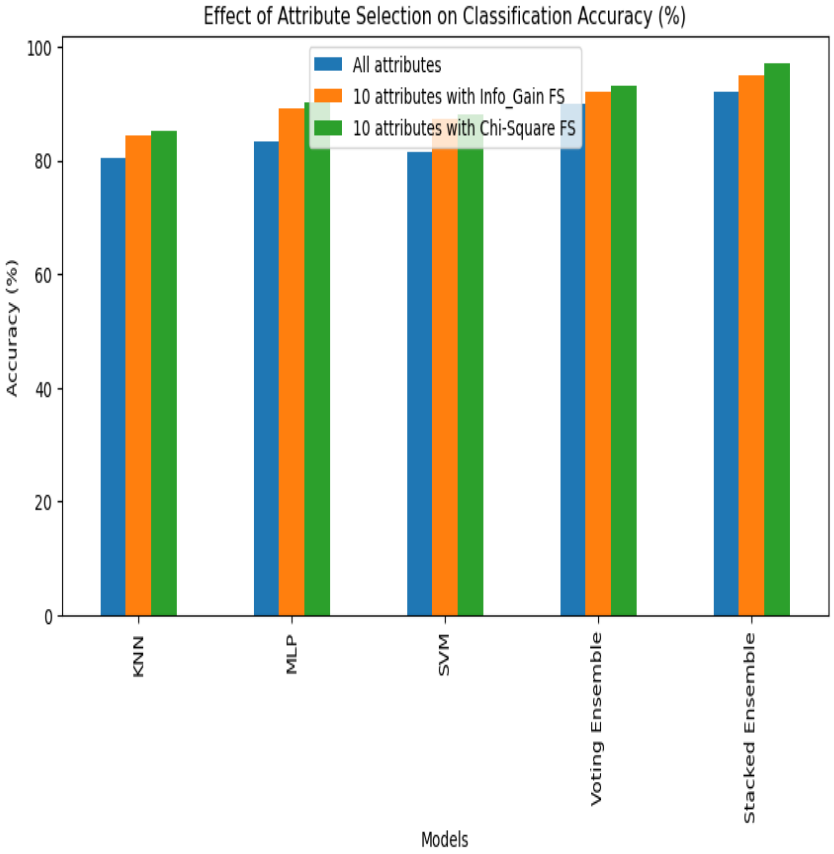


Figure 12 Effect of Attribute Selection on Classification Accuracy

4.4.2 Model Comparison

Following the experiments, the subsequent step was to compare the models and determine the best model. The experiments were conducted using three setups: one with all attributes, another with attributes selected using information gain feature selection, and the third with attributes selected using chi-square method. Various performance measures such as accuracy, Sensitivity (TP Rate), Specificity (TN Rate), Precision, F-Measure (F-1 score), and False Alarm Rate (FAR) were used to compare the models.

The stacked ensemble classifier achieved the highest classification accuracy of 97% when implemented on selected attributes from chi-square feature selection. It outperformed the voting ensemble, KNN, MLP, and SVM classifiers in predicting cases of coronary heart disease. The stacked ensemble model also had the highest TP Rate, TN Rate, Precision, F-Measure values, and the lowest false alarm rate of 3.2%. On the other hand, the K-nearest

neighbor classifier had the lowest accuracy of 80.42% and the highest false alarm rate of 93%.

All models performed well in terms of TP Rate and TN Rate, with minor variations in performance. Similarly, the Precision and F-1 Score were quite high for all models, indicating their ability to retrieve relevant values for each class. The False Alarm Rate (FAR) was also evaluated, and all models showed remarkable performance with minimal differences.

In conclusion, the stacked ensemble classifier implemented on selected attributes from chi-square feature selection method was chosen as the best predictive model. This model achieved the highest accuracy, TP Rate, TN Rate, Precision, F-Measure values, and the lowest false alarm rate. The stacked ensemble approach, which combines multiple machine learning algorithms, showed superiority in predicting cases of coronary heart disease and filtering out biases associated with a particular learning set.

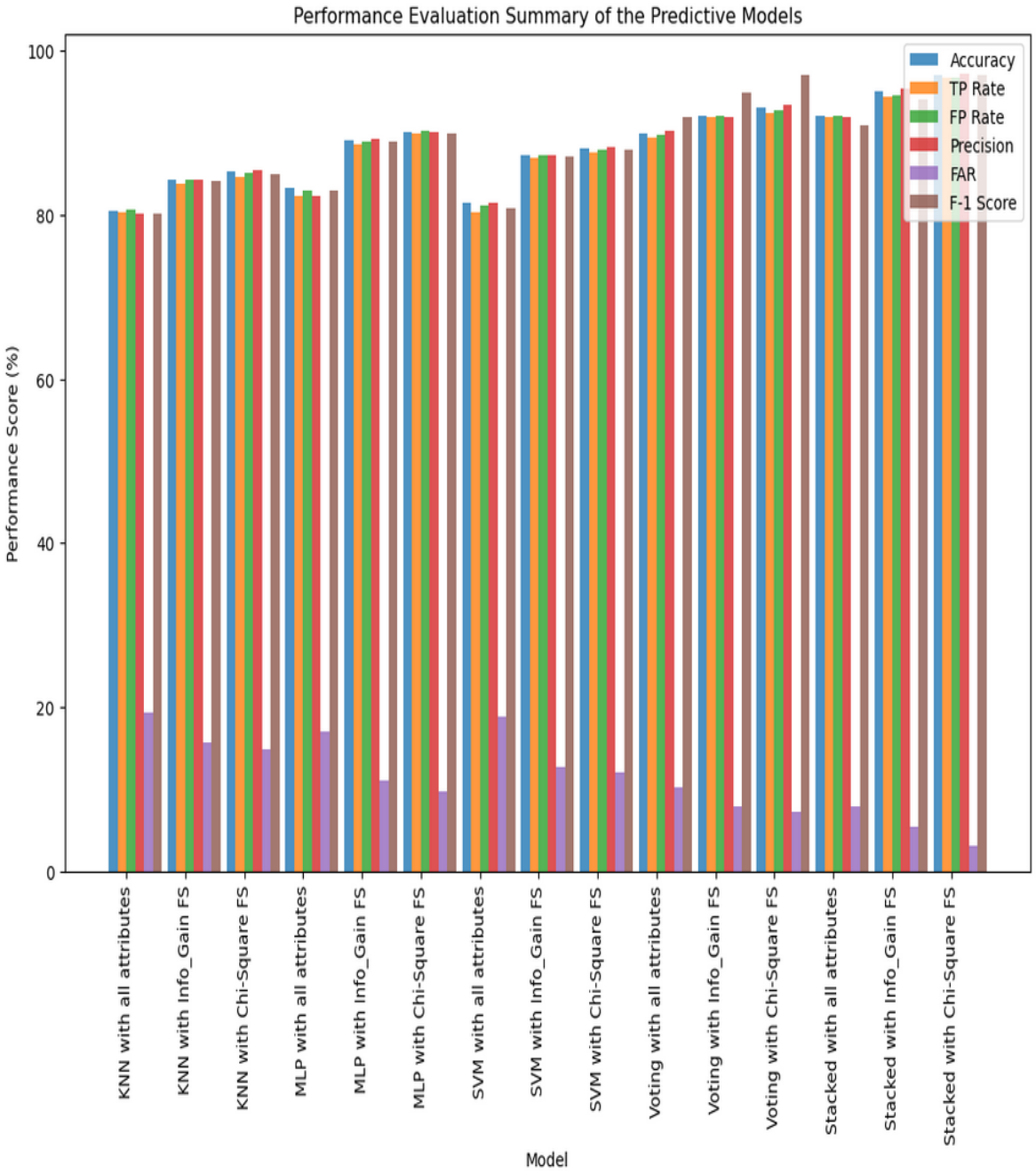


Figure 13 Performance Evaluation Summary of the Predictive Models

4.4.3 Comparison with other research works

Recent research have focused on comparing ensemble models with other approaches in various domains. These studies aim to assess the effectiveness and performance of ensemble models in comparison to individual machine learning algorithms or other ensemble methods. The findings provide valuable insights into the strengths and limitations of ensemble models and their potential for improving predictive accuracy and decision-making.

Comparing the thesis performance with other research can be found in the table below:

Table 14 Comparison with other works

Model	**Accuracy (%)**	**TP Rate (%)**	**FP Rate (%)**
Bashir *et al*., (2019) – NNE	80.93	82.9	78.5
Bashir *et al* (2021) – Bagging	74.61	81.1	66.9
Mao *et al*., (2016) – Adaboost	80.12	71.0	87.3
Thesis (1) - Voting with all attributes	90.02	89	89.8
Thesis (2) - Voting with Info_Gain FS	92.02	91.9	92.1
Thesis (3) - Voting with Chi-Square FS	93.02	92.4	92.7
Thesis (4) - Stacked with all attributes	92.02	91.9	92.1
Thesis (5) - Stacked with Info_Gain FS	95.01	94.4	94.6
Thesis (6) Stacked with Chi-Square FS	97.00	96.7	96.8

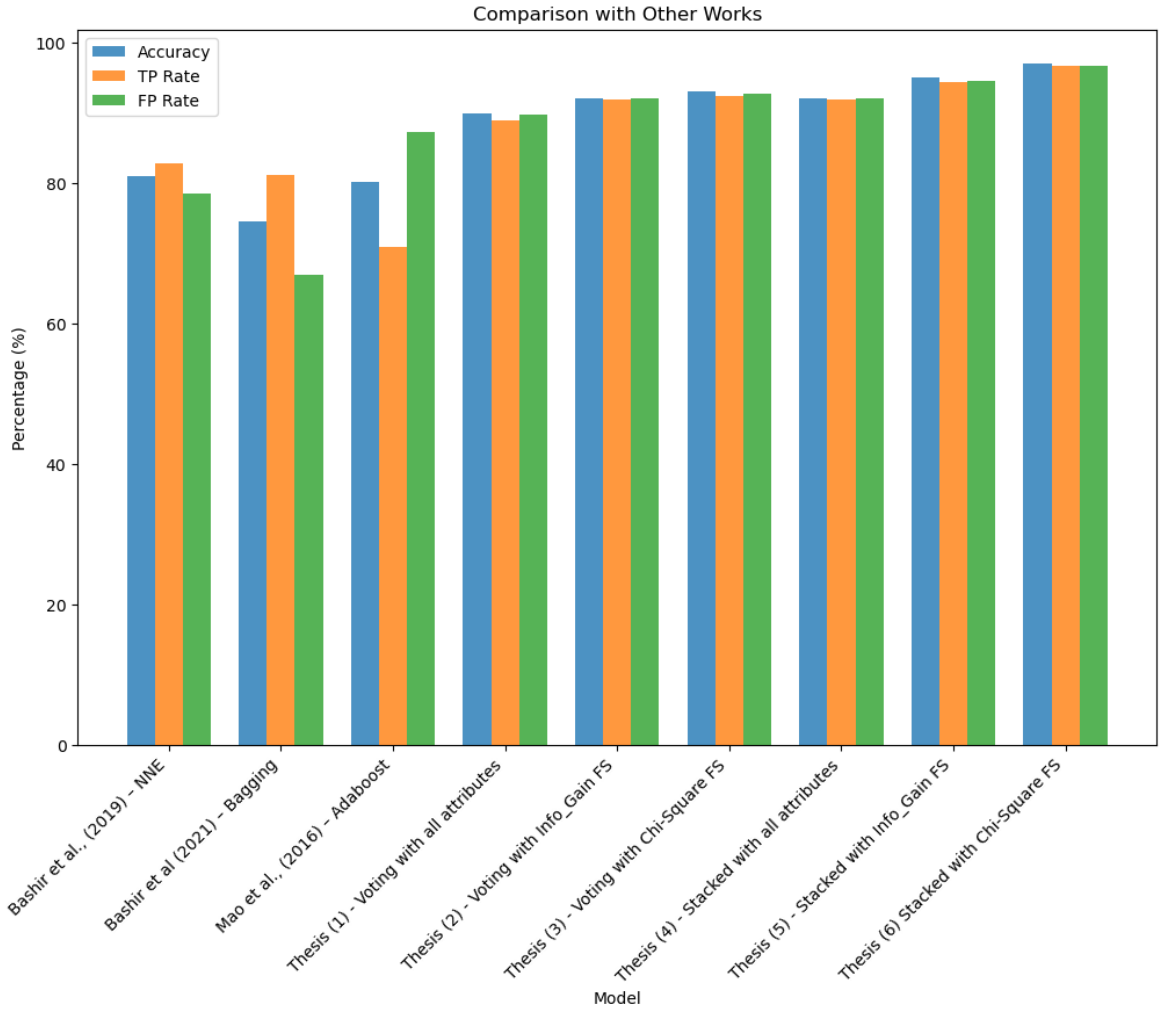


Figure 14 Comparison with other works

5 CONCLUSION

The main objective of this study was to develop reliable and accurate predictive models for detecting coronary heart disease using ensemble machine learning techniques. The dataset used in this study was obtained from UCI and consisted of 920 instances. Prior to modeling, the dataset was preprocessed and reduced to 802 instances.

Three supervised machine learning algorithms, namely K-Nearest Neighbor, Multilayer Perception, and Support Vector Machine, were used to build individual models. Additionally, ensemble voting and stacked generalization models were developed using the three algorithms as base models, with the latter using Logistic Regression as the Meta classifier. The models were implemented in Python, and their performances were evaluated using standard metrics such as accuracy, TP Rate, FP Rate, FAR, and F-measure. 10-Fold Cross Validation was employed to randomly sample the training and test data samples.

All fifteen models performed well in predicting coronary heart disease cases. However, the most effective model was found to be the stacked ensemble classifier implemented on selected attributes from chi-square, with a classification accuracy of 97%. From the original set of 13 attributes, the 10 most relevant attributes in predicting coronary heart disease were selected for the model.

Coronary heart disease is a serious and potentially life-threatening condition, and misdiagnosis can have serious consequences such as cardiac arrest, stroke, heart failure, and even death. Although the best model selected for predicting coronary heart disease achieved a classification accuracy of 97%, there is still room for improvement in order to reduce the misclassification rate of 3%.

The resulting model can serve as an assistant tool for cardiologists to improve their diagnosis of coronary heart disease and has high specificity rate, making it useful for less experienced cardiologists to identify patients who require further analysis and action by experienced cardiologists. Overall, this study highlights the potential of machine learning techniques in the medical field for improving disease diagnosis and treatment.

This thesis highlights the potential of machine learning techniques for predicting coronary heart disease and suggests that the resulting models are worth considering for clinical testing. Furthermore, the study shows that using feature selection techniques and stacked ensemble learning can improve the classification accuracy.

After developing the predictive models for CHD in this thesis, a better understanding of the relationship between attributes relevant to CHD was proposed. These models can be integrated into existing Health Information Systems (HIS), which capture and manage clinical information. The CHD diagnostic models can then be used to improve the real-time assessment of clinical information affecting patients in remote locations. It is recommended that continual assessment of CHD attributes be made to increase the number of relevant information for creating improved diagnostic models using the proposed methods of feature selection and machine learning algorithms. However, missing values, noisy data, inconsistencies, and outliers presented a challenge in the data mining process. Therefore, statistical and machine learning approaches should be used to control the quality of the data.

As a potential future work, it is recommended to conduct additional experiments using larger datasets and various algorithms in order to improve the classification accuracy for predicting CHD. It is also suggested to develop a model that can incorporate other types of heart related diseases to provide a more comprehensive diagnosis.

BIBLIOGRAPHY

[1] Abushariah, M., Alqudah, A., Adwan, O. and Yousef, R., (2014). Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) Approaches. Journal of Software Engineering and Applications, Volume 7, pp. 10551064.

[2] Akinyokun, C. O., Obot, O. U. and Uzoka, F. M. E. (2009). Application of neuro-fuzzy technology in medical diagnosis: Case study of heart failure. *IFMBE Proceedings*, *25*(12), pp. 301–304

[3] Amma, B. (2012). Cardiovascular disease prediction system using genetic algorithm and neural network. *2012 International Conference on Computing, Communication and Applications, ICCCA 2012*.

[4] Almuallim, H. and Dietterich, T. G. (1991) Learning with many irrelevant features," in *Proc. of the Ninth National Conference on Artificial Intelligence*, MIT Press, pp. 547-552.

[5] Anbarasi, M. Anupriya, E. and Iyengar N. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, *International Journal of Engineering Science and Technology*, vol.2 no.10, pp. 5370 – 5376.

[6] Alsalamah, M. (2017). *Heart Diseases Diagnosis Using Artificial Neural Networks*. Retrieved from https://core.ac.uk/download/pdf/156850229.pdf [Accessed October 12 2018].

[7] Ambekar, S. and Phalnikar, R. (2018). Disease Prediction by using Machine Learning. *International journal of computer engineering and applications*, Volume XII, special issue, pp. 1-6.

[8] ALPAYDIN, Ethem. *Introduction to machine learning*. Third edition. Cambridge, Massachusetts: The MIT Press, [2014], 1 online resource (xxii, 613 pages). Adaptive computation and machine learning. ISBN 9780262325745. Also available from: https://proxy.k.utb.cz/login?url=http://ieeeplore.ieee.org/servlet/opac?bknumber=6895440

[9] Begun, S. (2009). A Case-Based Reasoning System for the diagnosis of individual sensitivity to stress in psychophysiology, Ph.D Thesis, School of Innovation, Design and Engineering, Malardelen University, Sweden. Thesis No 102.

[10] Bergstra, J., Casagrande, N. Erhan, D. and B. K´egl, B. (2006) Aggregate features and AdaBoost for music classification. *Machine Learning*, 65(2-3), pp. 473–484.

[11] Benoît, F., Heeswijk, M., Miche, Y., Verleysen, M. and Lendasse, A (2013). Feature selection for nonlinear models with extreme learning machines, *Neurocomputin*g, vol. 102, pp. 111–124

[12] Bhatia, S, Prakash, P. and Pillai, G. N. (2008). SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features," Proceedings of the *World Congress on Engineering and Computer Science*, San Francisco, California, pp. 22-24

[13] Bharati, M. R. (2014). Data Mining Techniques and Applications. *Indian Journal of Computer Science and Engineering,* vol. 1 (4), pp. 301-305.

[14] Blum, A. L. and Langley, P. (1997). Selection of Relevant Features and Examples in Machine Learning, *Artificial Intelligence on Relevance*, vol. 97, pp. 245-271.

[15] Blum, A. L. and Rivest, R. L. (1998). Training a 3-node neural networks is NP-complete, Neural Networks, vol. 5, pp. 117-127.

[16] Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, vol. 34, no. 3, pp 483-519.

[17] Bradley, P. S. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines," in: Proc. 15th International Conference on Machine Learning (ICML), Madison, Wisconsin, USA, Morgan Kaufmann, pp. 82–90

[18] Breiman, L. (1996d). Bagging predictors. *Machine Learning*, 24(2), pp. 123–140.

Carpenter, G. A., Grossberg, S., and Rosen, D. B. (1991). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*(6), pp. 759–771.

[19] Caruana, R. A. and Freitag, D. (1994). Greedy Attribute Selection. in Proc. of the 11th Int. Conf. on Machine Learning, New Brunswick, NJ, Morgan Kaufmann, pp. 28-36.

[20] Centers for Disease Control and Prevention (CDCP) (2017), Heart Disease Facts. National Vital Statistics Reports, https://www.cdc.gov/heartdisease/facts.htm [Accessed February 3, 2018].

[21] Chan, P. K., Fan, W., Prodromidis, A. L. and Stolfo. S. J. (1999) Distributed datamin ing in credit card fraud detection. *IEEE Intelligent Systems*, 14(6), pp. 67–74.

[22] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), pp. 1–58

[23] Chaurasia V, and Pal, S. (2013), Data mining approach to detect heart disease. *International Journal of Advanced Computer Science and Information Technology*, Vol. 2, No. 4, pp. 56-66

[24] Choras, M., Bhanu, B., Chen, H., Champod, C., Komatsu, N., Nakano, M., and Liu, Z. (2009). Ensemble Learning. Encyclopedia of Biometrics, pp. 270–273.

[25] Comak, E. and Arslan, A., (2010). A Biomedical Decision Support System Using LS-SVM Classifier with an Efficient and New Parameter Regularization Procedure for Diagnosis of Heart Valve Diseases. *Journal of Medical Systems,* Volume 36, p. 549 – 556.

[26] Corona, I., Giacinto, G., Mazzariello, C., Roli, F. and Sansone C. (2009). Information fusion for computer security: State of the art and open issues. *Information Fusion*, 10(4), pp 274–284.

[27] Cortizo, J. C. and Giraldez, I. (2006). Multi Criteria Wrapper Improvements to Naive Bayes Learning," LNCS, vol. 4224, pp. 419–427

[28] Cui, Z., Yang, C. and Sanyal, S., (2012). Training artificial neural networks using APPM. *International Journal of wireless and mobile computing,* 5(2), pp. 168-174.

[29] Das, R., Turkoglu, I. and Al, E. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Journal of expert system with applications,* vol. 93, pp. 7675–7680.

[30] Dash, M. and Liu, H. (1997). Feature Selection for Classification, Intelligent Data Analysis, Elsevier, pp. 131-156

[31] David Opitz, D. and Maclin, R (2009). Popular Ensemble Methods: An Empirical Study, *Journal of Artificial Intelligence Research* 11, pp. 169-198.

[32] Deepali, C. (2014). Diagnosis of Heart Disease Using Data Mining Algorithm. *International Journal of Computer Science and Information Technologies (IJCSIT), 5*(2), pp. 1678-1680. Retrieved from www.ijcsit.com [Accessed September 18, 2018].

[33] Devi, A. D. (2015). *Enhanced Prediction of Heart Disease by Genetic Algorithm and RB network*. *2*(2), pp 271–276.

[34] Devi, S., Krishnapriya, S., and Kalita, D. (2016). Prediction of Heart Disease using Data Mining Techniques, https://doi.org/10.17485/ijst/2016/v9i39/102078. [Accessed October, 18 2018]

[35] Djam, X. Y., Wajiga, G. M., Kimbi Y. H. and Blamah, N. V. (2011) "A Fuzzy Expert System for the management of malaria", International Journal of Pure and Applied Sciences and Technology, pp.84-108.

[36] Doak, J. (1992). An Evaluation of Feature Selection Methods and Their Application to Computer Security," Technical Report, Davis CA: University of California, Department of Computer Science.

[37] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1),

[38]. Friedman, J., Hastie, H., and Tibshirani R. (2000). Additive logistic regression: A statistical view of boosting (with discussions). *Annals of Statistics*, 28(2):337– 407,

[39]. Gayathri, P. and Jaisankar, N. (2013). Comprehensive study of heart disease diagnosis using data mining and soft computing techniques. *International Journal of Engineering and Technology*, Vol. 5, No. 3, pp. 2947-2957.

[40]. Genders, T., Steyerberg, E., Hunink, M. and Nieman, K., (2012). Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. *British Medical Journal*, Volume 344, pp. 1-13.

[41]. Ghumbre, S., Patil, C. and Ghatol, A. (2011). Heart disease diagnosis using Support Vector Machine," *International Conference on Computer Science and Information Technology*, Pattaya, pp. 84-88.

[42]. Giacinto, G, Roli, F. and Didaci. L. (2003) Fusion of multiple classifiers for intrusion detection in computer networks. *Pattern Recognition Letters*, 24(12): 1795–1803.

[43]. Giacinto, G., Perdisci, R., Rio, M. D. and Roli, F. (2008). Intrusion detection in computer networks by a modular ensemble of one-class classifiers. *Information Fusion*, 9(1):69–82

[44]. Awad, M., Khanna, R. (2015). Machine Learning. In: Efficient Learning Machines. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_1

[45]. Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, Journal of Machine Learning Research, vol. 3, pp. 1157-1182.

[46]. Gudadhe, M., Wankhade, K. and S. Dongre, (2010) "Decision support system for heart disease based on support vector machine and artificial neural network", In proceedings of *IEEE International Conference on Computer and Communication Technology* (ICCCT), pp. 741–745

[47]. Guru, N., Dahiya, A. and Rajpal, N. (2007). Decision Support System for Heart Disease Diagnosis Using Neural Network, *Delhi Business Review*, vol. 8, No. 1,

[48]. Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning, in *Proc. of the 17th International Conference on Machine Learning*, pp. 359-366.

[49]. Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. Second Edition, *Morgan Kaufmann Publishers*, San Francisco

[50]. Hannan, S. A., Manza, R. R. and Ramteke, R. J., (2010). Generalized Regression Neural Network and Radial Basis Function for Heart Disease Diagnosis. *International Journal of Computer Applications,* 7(13), p. 7–13.

[51]. Hansen, L. K., and Salamon, P. (1990). Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. https://doi.org/10.1109/34.58871

[52]. Hedeshi, N. and Abadeh, M., (2014). Coronary Artery Disease Detection Using a Fuzzy-Boosting PSO Approach. *Computational Intelligence and Neuroscience,* (6).

[53]. Heller, R. F., Chinn, S., Tunstall, P. and Rose, G., (1984). How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. *British Medical Journal (Clinical Research Edition),* 12 May, 288 (6428), pp. 1409-1411.

[54] Helma, C., Gottmann, E. and Kramer, S., (2000). Knowledge discovery and data mining in toxicology.

[55] Heydari, S. T., Ayatollahi, S. M. & Zare, N., (2012). Comparison of Artificial Neural Networks with Logistic Regression for Detection of Obesity. *Journal of Medical Systems,* 36(4), pp. 2449-2454

[56] Higuera, V., (2014). Healthline Media Overview of Basics of Heart Disease. Available at: http://www.healthline.com/health/heart-disease/types#Overview. Accessed November, 2017.

[57] Hongzong, S., Jatmiko, W. and Murni, A., (2007). Support vector machines classification for discriminating coronary heart disease patients from non-coronary heart disease. *West Indian Medical Journal,* 56(5), pp. 451 – 457.

[58] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

[59] Huang, H., Liu, J. and Wang, G. H., (2014). A new hierarchical method for inter-patient heartbeat classification using random projections and RR intervals. *Biomedical Engineering,* pp. 1-26.

[60] Jang, J. S. R. (1993). ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Transactions on Systems, Man and Cybernetics*, *23*(3), pp. 665–685. [Assessed September, 19 2016]

[61] John, G., Kohavi, R. and Pfleger, K. (1994). Irrelevant feature and the subset selection problem, In *Proc. of the Eleventh International Conference on Machine Learning*, Morgan Kaufmann, ML – 94, pp. 121-129

[62] Kahramanli, K. and Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Journal of Expert Systems with Applications*, vol. 35, pp. 82-89.

[63] Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition. In *Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition*.

[64] Karabulut, E. & Ibrikci, T., (2012). Effective Diagnosis of Coronary Artery Disease Using Rotation Forest Ensemble Method. *Journal of Medical Systems,* Volume 36, pp. 3011-3018.

[65] Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S., and Pintelas, P. (2007). Feature Selection for Regression Problems. *The 8th Hellenic European Research on Computer Mathematics & Its Applications*, HERCMA, pp. 20–22.

[66] Karpagachelvi, S., Arthanari, M. and Sivakumar, M., (2011). Classification of ECG Signals Using Extreme Learning Machine. *Computer and Information Science,* 4(1).

[67] Kearns M. and Valiant, L. G. (1989). Cryptographic limitations on learning Boolean formulae and finite automata. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, pages 433–444, Seattle, WA.

[68] Kim, Y. S., Street, W. N. and Menczer, F. (2002). Evolutionary model selection in unsupervised learning, *Intelligent Data Analysis*, vol. 6, no. 6, pp. 531–556.

[69] Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm, In *Proc. of the Tenth National Conference on Artificial Intelligence*, MIT Press, pp. 129-134.

[70] Kohavi, R. and John, G. H. (1997). Wrapper for Feature Subset Selection. Artificial Intelligence, *Elsevier*, vol. 97, no. 12, pp. 273-324

[71]. Miao, K. H., Miao, J. H., and Miao, G. J. (2016). *Diagnosing Coronary Heart Disease Using Ensemble Machine Learning*. *7*(10), 30–39.

[72]. Karabulut, E. & Ibrikci, T., (2012). Effective Diagnosis of Coronary Artery Disease Using Rotation Forest Ensemble Method. *Journal of Medical Systems,* Volume 36, pp. 3011-3018.

[73]. Obot, O. U. , Uzoka, F. M. , Akinyokun O. C. and Andy, J. J. (2013) Conventional and neuro-fuzzy framework for diagnosis and therapy of cardiovascular disease. Journal of *Bio-Algorithms and Med-Systems* Vol. 9, no. 3, pp. 115--133

[74]. OJEDA, Tony, Sean Patrick MURPHY, Benjamin BENGFORT, and Abhijit DASGUPTA. *Practical data science cookbook: 89 hands-on recipes to help you complete real-world data science projects in R and Python*. Birmingham: Packt Publishing, 2014, 380 pp. ISBN 9781783980246.

[75]. Otoom, A. F., Abdallah, E. E., Kilani, Y., Kefaye, A. and Ashour, M, (2015). Effective diagnosis and monitoring of heart disease. *International Journal of Software Engineering and Its Applications*, Vol. 9, No. 1, pp. 143-156.

[76]. Liu, H. and Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers,

[77]. Turing, M. A. (1950) Computing Machinery and Intelligence. Mind 49, pp 433-460.

[78]. Sayad, A. T., and Halkarnikar, P. P. (2014). Diagnosis of Heart Disease Using Neural Network. *International Journal of Advances in Science Engineering and Technology*, *2*(3), pp. 88–92.

[79]. Setiawan, N. A., Venkatachalam, P. A. and Fadzil, M. H. A. (2009). Rule selection for coronary artery disease diagnosis based on rough set. *International Journal of Recent Trends in Engineering,* 2(5), pp. 198-202.

[80]. Singh, N. and Jindal, S. (2018). Heart Disease Prediction System using Hybrid Technique of Data Mining Algorithms, vol. *4*(2), pp. 982–987.

[81] Narendra, P. and Fukunaga, K. (1997), "A Branch and Bound Algorithm for Feature Subset Selection. *IEEE Transactions on Computer*, vol. 26, no. 9, pp. 917-922. .

[82] Vapnik, V., Golowich, S. and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing, *Advances in Neural Information Processing Systems* 9, vol. 9, pp. 281-287.

[83]. Molina, L. C., Belanche, L. and Nebot, A. (2002) "Feature Selection Algorithms: A Survey and Experimental Evaluation," in *Proc. of ICDM,* pp. 306-313

[84]. Viola, P. and Jones, M. (2004). Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154.

[85]. Wang, W. and Zhou, Z. (2008). On multi-view active learning and the combination with semi supervised learning. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1152–1159, Helsinki, Finland,

[86] West, D., Dellana, S. and Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers and Operations Research*, 32(10), pp. 2543–2559

[87] WHO (2011). Fact Sheet: The Top Ten Causes of Death. World Health Organization. Geneva 941 WHO (2016). Fact Sheet: Cardiovascular Diseases.

[88] World Health Organization. Geneva Witten WHO, 2017. World Health Organization, Media Centre, cardiovascular diseases fact sheet webpage. http://www.who.int/mediacentre/factsheets/fs317/en/ Wilson, P. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories, *American Heart Association Journal*.

[89] Witten H. And Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Second Edition. *Morgan Kaufmann Publishers*, San Francisco

[90]. Mukhamediev, Ravil. (2015). Machine learning methods: An overview. CMNT. 19. 14-29.

[91]. Yan, W. and Xue, F. (2008). Jet engine gas path fault diagnosis using dynamic fusion of multiple classifiers. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1585– 1591, Hong Kong, China.

[92]. Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: a fast correlation-based filter solution. Proc. 20th Int‟l Conf. Machine Learning, pp. 856-863.

[93]. MILES, Matthew B., AM HUBERMAN, and Johnny SALDAÑA. *Qualitative data analysis: a methods sourcebook*. Fourth edition. Los Angeles: SAGE, [2020], xxi, 380 pp. ISBN 9781544371856.

[94]. DORSEY, Richard. *Data analytics*. [Create Space Independent Publishing Platform], [2017], 67 pp. ISBN 9781547089291.

[95]. Zhou, Z., Jiang, Y., Yang, Y. and Chen, S. (2002a). Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine*, 24(1), pp. 25–36.

[96]National Heart Lung and Blood Institute (2008). What is Echocardiography?.Available at http://www.nhlbi.nih.gov/health/dci/Diseases/echo/echo_whatis.html [Accessed March, 17 2017]

[97]National Heart, L. A. B. I., 2016. What Is Coronary Heart Disease? Available at: http://www.nhlbi.nih.gov/health/health-topics/topics/cad.[Accessed,February12 2018]

[98] Miao, K. H., Miao, J. H., and Miao, G. J. (2016). *Diagnosing Coronary Heart Disease Using Ensemble Machine Learning*. *7*(10), 30–39.

[99]Kuulasmaa, K., Tunstall-Pedoe, H., Dobson, A., Fortmann, S., Sana, S., Tolonen, H., Eans, A., Ferrario, M. and Tuomilehto, J. (2000). Estimation of Contribution of Changes in Classic Risk Factors to Trends in Coronary Event Rates across *WHO MONICA Project Populations. Lancet* 355, pp. 675–687

[100] Bisong, E. (2019). The Multilayer Perceptron (MLP). In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_31

[101] Muhammad, Y., Tahir, M., Hayat, M. *et al.* Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Sci Rep* **10**, 19747 (2020). https://doi.org/10.1038/s41598-020-76635-9

[102]RahulKumar.,https://medium.com/@rahul.apiit23/decode-confusion-matrix-bb554c299d01 bb554c299d01

[103] Silva, R., & Wilcox, S. (2019). Feature evaluation and selection for condition monitoring using a self-organizing map and spatial statistics. AI EDAM, 33(1), 1-10. doi:10.1017/S0890060417000518

[104] Ralapanawa U, Sivakanesan R. Epidemiology and the Magnitude of Coronary Artery Disease and Acute Coronary Syndrome: A Narrative Review. J Epidemiol Glob Health. 2021 Jun; 11(2):169-177. doi: 10.2991/jegh.k.201217.001. Epub 2021 Jan 7. PMID: 33605111; PMCID: PMC8242111.

List of abbreviations

HD: Coronary Heart Disease - A condition where the blood vessels that supply the heart with oxygen and nutrients become narrowed or blocked, leading to various heart-related complications.

KNN: K-Nearest Neighbor - A supervised machine learning algorithm used for classification and regression tasks. It determines the class of a sample by identifying the majority class among its k nearest neighbors in the feature space.

MLP: Multilayer Perceptron - A type of artificial neural network composed of multiple layers of interconnected nodes (neurons). It is commonly used for supervised learning tasks, including classification and regression.

SVM: Support Vector Machine - A supervised machine learning algorithm that constructs a hyperplane or set of hyperplanes to separate data into different classes. It is often used for classification tasks.

TP: True Positive - The number of correctly predicted positive instances in binary classification.

TN: True Negative - The number of correctly predicted negative instances in binary classification.

FP: False Positive - The number of incorrectly predicted positive instances in binary classification.

FN: False Negative - The number of incorrectly predicted negative instances in binary classification.

FAR: False Alarm Rate - The proportion of falsely predicted positive instances in binary classification.

F-Measure: Also known as F1 Score, it is a measure that combines precision and recall to evaluate the overall performance of a classification model.

HIS: Health Information Systems - Computer-based systems used to capture, store, manage, and exchange health-related information for healthcare providers.

UCI: University of California, Irvine - Refers to the dataset source, which is commonly used in machine learning research and experimentation.

FS: Feature Selection - A process of selecting a subset of relevant features from a larger set of available features to improve model performance and reduce computational complexity.

LIST OF FIGURES

LIST OF TABLES

APPENDICES