International Journal of Research and Applied Technology

3(2) (2023) 267-282
Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Machine Learning Approaches for Heart Disease Detection: A Comprehensive Review

Hanan A. Taher[1*], Adnan M. Abdulazeez[2]

[1]Information Technology Department, Duhok Technical College, Polytechnic University, Iraq

[2]Technical College of Engineering, Duhok Technical College, Polytechnic University, Iraq

Email: * hanan.taher@dpu.edu.com

Abstract. This paper presents a comprehensive review of the application of machine learning algorithms in the early detection of heart disease. Heart disease remains a leading global health concern, necessitating efficient and accurate diagnostic methods. Machine learning has emerged as a promising approach, offering the potential to enhance diagnostic accuracy and reduce the time required for assessments. This review begins by elucidating the fundamentals of machine learning and provides concise explanations of the most prevalent algorithms employed in heart disease detection. It subsequently examines noteworthy research efforts that have harnessed machine learning techniques for heart disease diagnosis. A detailed tabular comparison of these studies is also presented, highlighting the strengths and weaknesses of various algorithms and methodologies. This survey underscores the significant strides made in leveraging machine learning for early heart disease detection and emphasizes the ongoing need for further research to enhance its clinical applicability and efficacy.

Keywords: Machine Learning, Classification Techniques, Supervised Learning, Naïve Bayes, Support Vector Machine, Heart Disease, Decision Trees, K- Nearest Neighbor, Random Forest

ARTICLE INFO:
Submitted/Received 1 May 2023
First revised 12 Jul 2023
Accepted 18 Oct 2023
First available online 08 Dec 2023
Publication date 25 Dec 2023

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

1. INTRODUCTION

Cardiovascular diseases (CVDs) persist as the foremost cause of mortality globally, exacting a profound toll on public health and healthcare systems worldwide [1]. Among the diverse spectrum of CVDs, heart disease, which encompasses conditions such as coronary artery disease, heart failure, and arrhythmias, remains a particularly formidable adversary [2]. The imperative for early detection and intervention in heart disease cannot be overstated, as it directly impacts patient outcomes, healthcare costs, and societal well-being [3].

Machine learning, an advanced segment of artificial intelligence, has shown remarkable capabilities in various fields, including the assessment of facial attractiveness and disease prediction. In facial attractiveness prediction, machine learning algorithms are trained using vast datasets of facial images, each rated for attractiveness by human observers [4]. These algorithms then learn to identify patterns and features that correlate with perceived attractiveness [5]. Deep Neural Networks is used for the task of facial attractiveness assessment by applying knowledge gained from one domain and applies it to a related but different domain [6, 7]. This approach is particularly beneficial in scenarios with limited data. This technology not only aids in understanding human perceptions of beauty but also finds applications in cosmetic surgery, advertising, and social media filters. In the realm of disease prediction, machine learning analyzes vast medical data to predict diseases, like early breast cancer detection [8] or echocardiograms for heart disease prediction.

In recent years, the convergence of healthcare data proliferation, computational prowess, and machine learning (ML) methodologies has reshaped the landscape of heart disease detection [9]. The utility of ML algorithms, driven by their aptitude for deciphering intricate patterns within large and heterogeneous datasets, holds immense promise in revolutionizing how we approach the identification, stratification, and management of heart disease [10]. These algorithms can assimilate multifaceted data sources, encompassing electronic health records, medical imaging, genetic markers, and lifestyle variables, to unveil latent associations and prognostic insights that might elude human cognition [11].

This review paper undertakes a thorough investigation of the diverse functions of machine learning in the domain of detecting heart disease. Leveraging a wealth of empirical research, clinical studies, and technological innovations, our objective is to offer a comprehensive overview of the current state of ML-powered solutions for heart disease diagnosis. We delve into the underlying methodologies and techniques underpinning ML-driven cardiac risk assessment, elucidate notable achievements in terms of diagnostic accuracy and prediction, and dissect the associated challenges and limitations. Furthermore, we endeavor to delineate the transformative impact of ML on patient care pathways, healthcare resource allocation, and population health management, all while emphasizing the necessity of collaborative endeavors between clinicians, data scientists, and policymakers in the pursuit of a data-driven and patient-centric approach to tackling this pervasive global health challenge. As we traverse the terrain of ML in heart disease detection, we highlight both the successes that propel us toward a brighter future in cardiac care and the obstacles that demand innovative solutions. This review is a testament to the remarkable strides made in leveraging ML for cardiovascular health and a clarion call for continued research and development to harness the full potential of these transformative technologies.

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

The Cleveland Heart Disease dataset, derived from the UCI Machine Learning Repository, serves as a cornerstone in this transformative journey. This dataset, comprising 303 instances and 14 essential attributes, has served as a pivotal resource for researchers and clinicians seeking to develop and validate ML-driven solutions for cardiac risk assessment [12].

2. IMPORTANCE OF EARLY HEART DISEASE PREDICTION

Early heart disease prediction is crucial for preventing the development and progression of heart disease, improving health outcomes, reducing healthcare costs, and enhancing the overall quality of life for individuals at risk. It also has broader societal and public health implications by reducing the burden of heart disease on healthcare systems and promoting healthier communities. Early heart disease prediction is of paramount importance for several reasons:

i) Prevention: Early prediction of heart disease allows individuals to take proactive measures to reduce their risk factors and adopt a healthier lifestyle. Lifestyle changes, such as improving diet, increasing physical activity, quitting smoking, and managing stress, can significantly reduce the risk of heart disease.

ii) Improved Health Outcomes: Identifying heart disease at an early stage can lead to more effective treatment and better health outcomes. Early intervention can prevent the progression of the disease and reduce the risk of complications such as heart attacks, strokes, and heart failure.

iii) Reduced Healthcare Costs: Detecting heart disease early can lead to cost savings in the healthcare system. Treating advanced stages of heart disease is often more expensive and requires more resources than managing risk factors and early-stage disease.

iv) Quality of Life: Early prediction and intervention can enhance the quality of life for individuals with heart disease. It can help them maintain their independence, stay active, and enjoy a higher quality of life without the limitations and symptoms associated with advanced heart disease.

v) Personalized Medicine: Predictive tools and technologies allow for more personalized treatment plans. Healthcare providers can tailor interventions based on an individual's specific risk factors and genetic predispositions, leading to more effective and targeted care.

vi) Reduction in Mortality: Early detection and intervention can significantly reduce the risk of mortality from heart disease. Identifying and managing risk factors can prevent heart attacks and other life-threatening events.

vii) Public Health Impact: Early heart disease prediction can have a broader public health impact by reducing the overall burden of heart disease on society. This can lead to a healthier population and a decrease in the strain on healthcare systems.

viii) Research Advancements: Early prediction also contributes to ongoing research into heart disease. It provides valuable data for understanding the progression of the disease and the effectiveness of various preventive measures and treatments.

ix) Patient Empowerment: When individuals are aware of their risk factors and the early signs of heart disease, they can take an active role in managing their health. This empowerment has the potential to result in enhanced adherence to treatment plans and the adoption of healthier lifestyle decisions.

x) Long-Term Planning: Early prediction allows individuals and their healthcare providers to engage in long-term planning. They can develop strategies to manage the disease over time, ensuring that the individual's health needs are met as they age.

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

3. THEORETICAL BACKGROUND

Supervised and unsupervised machine learning are two fundamental paradigms in the field of artificial intelligence and machine learning, each with its own unique characteristics and applications. Let's delve into the theoretical background of both approaches.

4. UNSUPERVISED MACHINE LEARNING

Unsupervised learning deals with unlabeled data, where the algorithm must find patterns, structures, or relationships within the data without any predefined target variable [13]. The primary objective is to discover hidden patterns or groupings within the data. Unsupervised learning often involves clustering, which groups similar data points together based on some similarity measure [14]. K-means clustering and hierarchical clustering are common clustering techniques.

Unsupervised learning can also be used for reducing the dimensionality of high-dimensional data. T-Distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA) are examples of dimensionality reduction techniques [15]. Certain unsupervised algorithms are designed to estimate the probability distribution of the data, and this capability can be valuable for tasks such as anomaly detection and generative modelling [16]. Gaussian Mixture Models (GMMs) and kernel density estimation fall into this category.

5. Unit of Analysis and Unit of Observation:

Supervised learning is a category of machine learning in which the algorithm is trained using a labeled dataset [17]. In this context, "labeled" means that each input data point is associated with a corresponding output or target value. The central objective of supervised learning is to acquire a mapping from input data to the desired output, allowing the algorithm to make predictions or classifications for new, unseen data [13]. Following is a summary of some machine learning algorithms.

5.1. Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a machine learning model that draws its inspiration from the architecture and operation of the human brain [18]. It is comprised of layers of interconnected nodes, known as neurons, which handle the processing and transmission of information. ANNs can be used for various tasks, including classification, regression, and pattern recognition. Training involves adjusting the weights of connections to minimize a loss function through techniques like backpropagation [19]. Deep Neural Networks (DNNs) are advanced Artificial Neural Networks characterized by their multiple hidden layers, which enable them to effectively learn and discern complex patterns in data, such as classifying hand-written digits in images. This architecture is key to their proficiency in handling intricate tasks in various fields, including image and pattern recognition [20]. Optimization is crucial in training Artificial Neural Networks (ANNs), involving algorithms like Gradient Descent to minimize loss functions [21]. It's essential for hyperparameter tuning, model selection, and preventing overfitting, ensuring efficient and effective performance [22]. Metaheuristic algorithms contribute to enhancing the performance of classification models in deep learning and machine learning by efficiently solving the feature selection problem [23]. They assist in identifying the most relevant features for binary classification, using various classifiers, datasets, and evaluation metrics. This approach helps in optimizing the feature subset, thereby improving the classification accuracy and model efficiency [24]. the Giant Trevally Optimizer

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

(GTO) exemplifies the innovative advancements in metaheuristic algorithms [25], showcasing their capability to efficiently tackle and resolve intricate optimization problems across diverse domains [26]. Another Optimization algorithm is NGO, a type of optimization algorithm used in machine learning and other computational fields. The concept behind the NGO algorithm is inspired by the hunting behavior of Northern Goshawks, a species of bird known for its efficient and strategic hunting skills [28].

5.2. K-Nearest Neighbour (KNN)

KNN is a simple and intuitive algorithm for classification and regression. It classifies data points by finding the K nearest neighbors in the training dataset and taking a majority vote (for classification) or averaging (for regression) [29]. The choice of K determines the smoothness and accuracy of the model, with smaller K values leading to more complex decision boundaries [30]. KNN can be sensitive to the scale of features and requires careful preprocessing. It is a lazy learner, meaning it does not build an explicit model during training, making it computationally inexpensive but potentially slow during prediction. KNN is widely used in the medical field for diagnosing diseases and conditions [31]. It helps in classifying patients based on similarity in symptoms and historical data, aiding in accurate disease identification [32]. In the realm of e-commerce and online services, KNN algorithms play a crucial role in creating recommender systems. These systems suggest products, movies, or music to users by finding similar items based on their past preferences and the preferences of other users with similar tastes [33, 34].

5.3. Naïve Bayes

Naïve Bayes is a probabilistic algorithm that utilizes Bayes' theorem and operates under the assumption of feature independence [35]. It is commonly used for text classification and spam filtering. It assumes that all features contribute independently to the class probability, which can be a limitation in some cases [36]. Despite its simplicity and naive assumption, Naïve Bayes often performs surprisingly well on various classification tasks. Naive Bayes can assist in medical diagnosis by analyzing patient data and symptoms to predict the likelihood of a particular disease or condition [37]. It can be used to detect the language of a given text or document, which is useful in multilingual applications [38].

5.4. Decision Tree Algorithm

Decision Trees are flexible machine learning models employed for tasks involving classification and regression [39]. They take the form of a tree-like structure in which every internal node signifies a decision based on a feature, and each leaf node signifies either a class or a numeric value [40]. Decision Trees are interpretable and can handle both categorical and numerical data. They can suffer from overfitting if the tree is too deep, but this can be mitigated with techniques like pruning. Random Forests and Gradient Boosted Trees are ensemble methods built on Decision Trees, offering improved performance and robustness. Banks and financial institutions use decision trees to assess credit risk by determining whether a loan applicant is likely to default based on factors such as income, credit history, and employment status [41]. Decision trees can be used to identify fraudulent transactions in financial systems by analyzing transaction attributes and flagging unusual or suspicious behaviour [42].

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

5.5. Random Forest

Random Forest is an ensemble learning technique that relies on Decision Trees. During training, it creates multiple decision trees and then aggregates their predictions to achieve more accurate and robust classifications or regressions [43]. Random Forests are known for their high predictive performance and versatility. Each tree is constructed from a bootstrapped subset of the training data. Random feature selection is applied at each split to reduce correlation between trees. Ensemble predictions are made by aggregating the results from individual trees. Random Forests are widely used in various domains, including healthcare, finance, and natural language processing, due to their ability to handle complex data [44]. Random Forest is used for species classification, land cover classification, and predicting environmental phenomena such as air quality or deforestation rates [45].

5.6. Support Vector Machine (SVM)

SVM is supervised learning algorithm employed for tasks involving classification and regression. It seeks to identify the optimal hyperplane that maximizes the margin between distinct classes within a high-dimensional feature space [46]. It can handle non-linear data by using various kernel functions, such as radial basis function (RBF) and polynomial kernels. SVM is known for its ability to generalize well and perform effectively in high dimensional spaces. Applications of SVM include image classification, text categorization, and bioinformatics [47]. Marketing and customer segmentation used SVMs to segment customers based on behavior, preferences, and demographics for targeted marketing campaigns [48].

6. LITERATURE SURVEY

In recent years, there have been notable advancements in early heart disease detection and management, primarily driven by the application of advanced computer algorithms, especially those rooted in Machine Learning (ML). ML algorithms have proven highly effective in forecasting various medical conditions, leading to a significant shift in research focus towards their utilization for early heart disease prediction. This section provides a comprehensive survey of the latest ML models designed for this purpose.

Senthilkumar et al. developed machine learning algorithms NB, DL, LR, GLM, GBT, RF, SVM, DT, HRFLM (RF+LM) to predict the presence of heart diseases in patients [49]. The dataset used was the Cleveland Heart Diseases and is collected from UCI machine learning repository which contains information on patients with heart disease. The dataset has 14 attributes and measured on 303 individuals. The proposed hybrid HRFLM approach combines the attributes of Random Forest (RF) and Linear Method (LM). HRFLM has demonstrated its effectiveness in predicting heart disease, achieving an impressive accuracy rate of 88.4%.

Kannan et al. [49] developed a machine learning model for predicting heart disease. To build this model, the Cleveland Heart Disease Dataset from the UCI Repository was utilized which contains 14 distinct parameters related to heart disease. Machine learning algorithms such as Support Vector Machine (SVM), Random Forest, Logistic Regression, and Stochastic gradient boosting have been used for the development of model. The methods have been used efficiently in the prediction of Heart disease. The results demonstrated that Logistic Regression offered superior accuracy (86%) compared to other machine learning methods.

Jianping et al. [51] have put forth an effective system for diagnosing heart disease, utilizing machine learning classification algorithms and the Cleveland heart disease data from UCI. The system was designed using a combination of machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression, Artificial Neural Network, K-nearest neighbor,

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Naïve Bayes, and Decision Tree algorithms. For solving feature selection problem, proposed novel fast conditional mutual information feature selection algorithm is used (FCMIM). The proposed system (FCMIM+SVM) has demonstrated remarkable accuracy rate 92.37% as compared to other features selection algorithms and classifiers.

Archana et al. [52] involved the calculation of accuracy for various machine learning algorithms aimed at predicting heart disease by using UCI repository dataset (which have 303 samples with 14 input feature) for training and testing. SVM, DT, LR, K-NN Machin learning algorithms are performed for training and testing. It concludes that K-NN is much efficient as compare to other algorithms with 87% accuracy.

Anna et al. [53] proposed the use of a PCA with chi-square (CHI) to improve the prediction of heart disease using machine learning models. In addition, dimensionality reduction is applied to improve the results by selecting features which achieved the best performance. Different classifiers have been investigated such as MPC, DT, GBT, LOG, RF, and NB. It was found that Chi-square and principal component analysis (CHI-PCA) with RF had the maximum performance, with 98.7% accuracy for for Cleveland dataset. The primary was to determine the most effective dimensionality reduction technique for heart disease prediction in terms of performance. As a result, CHI-PCA emerged as the most consistent and favored method in their analysis.

Yar et al. [54] created a system for the detection of heart disease using machine learning models. They conducted training and testing using both the complete set of features and a carefully selected subset of optimal features. The study analysis of the outcomes achieved by each feature selection algorithm in conjunction with various classifiers. Ten classification algorithms including, KNN, DT, RF, NB, SVM, AB, ET, GB, LR, and ANN, and the authors applied four distinct feature selection algorithms such as FCBF, mRMR, LASSO, and Relief are used. The paper utilized the 10-Fold cross-validation method to validate the performance results of the classification models mentioned earlier. Additionally, P-values and Chi-square are also computed for the ET classifier in conjunction with each feature selection technique. the ET classifer with the relief feature selection algorithm performs excellently with accuracy rate 94.41%.

Rohit et al. [55] applied deep learning and different machine learning algorithms to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. Machine learning algorithms performed better in this analysis when data pre-processing is applied. Linear regression, K neighbors, Support Vector Machine, Random Forest, and Decision Tree algorithm are utilized and various promising results are achieved. confusion matrix, precision, specificity, sensitivity, and F1 score methods are used for validating results. Lasso model has been used for feature selection which gives improvements in the algorithms accuracy. The maximum accuracy achieved by KNeighbors (84.8%). Feature selection is done and also the outliers are handled using the Isolation Forest.

M.kavita et al. [56] devised a hybrid model that utilizes both decision tree and random forest algorithms. This combined model relies on the probability outputs generated by the random forest. The probabilities originating from the random forest are combined with the training data and input into the decision tree algorithm. In a corresponding manner, the probabilities generated by the decision tree are pinpointed and integrated into the test data. Finally, values are predicted. The proposed study used the Cleveland heart disease dataset and three machine learning algorithms: Decision Tree, Random Forest, and Hybrid model (Hybrid of random forest and decision tree). The experimental findings indicate that the heart

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

disease prediction model achieved its highest accuracy, reaching 88.7%, when utilizing the hybrid model.

Abdul Saboor et al. [57] performed a comparative study on different machine learning algorithms with the aim of developing a heart disease prediction model with improved and enhanced accuracy. The algorithms investigated were ET, LR, AB, CART, MNB, LDA, SVM, XGB and RF. From the comparison of different classifiers, the study concludes that XGB and ET classifiers demonstrate generally strong accuracy. However, SVM shows the best accuracy in tuning the hyper parameters and achieved an accuracy of 96.72%.

Niloy et al. [58] conducted an evaluation of various machine learning techniques for heart prediction. Their study encompassed six distinct machine learning models, namely support vector machine, random forest, logistic regression, K-nearest neighbor, decision tree, and Naïve Bayes. In the conducted research, the Cleveland heart disease dataset was utilized, and three different approaches for feature selection were employed: chi-square, ANOVA, and mutual information. Among these methods, Random Forest demonstrated the most promising performance, achieving an accuracy rate of 94.51%.

7. MEDICAL DATASET USED

All researches covered in this survey exclusively employ the Cleveland heart disease dataset have all utilized the Cleveland heart disease data set which is retrieved from the UCI machine learning repository (UCI Machine Learning Repository, 2023). It comprises a real dataset of 303 examples of data with 14 various attributes (13 predictors; 1 class).

8. FEATURE SELECTION

Attributes within a dataset are essentially the properties or characteristics that serve as the basis for determining whether an individual is afflicted with a particular disease. These attributes encompass a wide range of factors, including heart rate, types of chest pain, gender, age, the presence of exercise-induced angina, and numerous other features. Table 1 provides detailed descriptions of the 14 features of the dataset. Feature selection is a vital component of the machine learning process because there are instances where datasets consist of numerous unimportant features, and these can have a detrimental impact on the accuracy of algorithms (Zhou et al., 2022) . It used diverse feature ranking techniques to determine and rank the most significant feature based on its relevance [59]. Integrating clinically important features with those extracted through deep learning methods is highlighted, suggesting a preference for a hybrid approach that leverages both traditional and advanced techniques in feature extraction and analysis. As a case in point, combining hand-crafted features with features derived from deep networks is valued to enhance the robustness of peripapillary atrophy detection [60].

9. DATA EXPLORATION

This step serves the purpose of visually exploring the dataset, aiming to gain an understanding of the various biological parameters contained within it, as illustrated in Figure 1. The dataset comprises records from 303 patients, with 165 of them having heart disease. The dataset features a gender distribution ratio of 32:69 between female and male patients. Likewise, patients with Types 0, 1, and 2 chest pain are present in the ratios of 143:50:87:23, respectively. Patients with blood sugar levels up to 120 mg/dL are found in the ratio of 15:86, and those with and without Exercise-Induced Angina are in the ratio of 3:68. The age of 54 stands out as having the highest frequency of heart disease among the age group spanning from 29 to 77. The dataset's target class labels were distributed as follows: Label 0, indicating "no risk of heart

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

disease," comprised 164 instances, making up 54% of the dataset, while Label 1, indicating "risk of heart disease," comprised 139 instances, accounting for 46% of the dataset, as illustrated in Figure 2.

Table 1. Dataset Attributes Description

Name	Definition	Type	Description
Age	Age	Integer	Age in years
Sex	Sex	Categorical	0=female 1=male
Chest pain	Cp	Categorical	Chest pain type: 1=typical angina 2=atypical angina 3=non-anginal pain
Resting blood pressure	trestbps	Integer	Resting blood pressure (mm/Hg)
Serum cholesterol	chol	Integer	Cholesterol (mg/dl)
Fasting blood sugar	Fbs	Categorical	Fasting blood sugar > 120 (mg/dl): 1=true 0=false
Rest electrocardiograph	restecg	Categorical	Resting electrocardiographic result: 0=normal 1=having ST-T abnormality 2= probable left ventricular Hypertrophy
MaxHeart rate	thalach	Integer	Maximum heart rate achieved
Exercise-Induced angina	exang	Categorical	Exercise-induced angina: 1=yes 0=no
ST depression	oldpeak	Integer	ST depression induced by exercise relative to rest
Slope	slop	Categorical	Peak exercise slope segment: 1=up sloping 2=flat 3=down sloping
No. of vessels	Ca	Integer	Number of major vessels colored by fluoroscopy that ranges from 0-3
Thalassemia	thal	Categorical	Heart rate: 3=normal 6=fixed defect 7=reversible defect
Class	Num (class attribute)	Integer	Diagnosis classes: 0=healthy 1=possible heart disease

10.DATA PREPROCESSING

Data preprocessing plays a pivotal role in improving the performance and robustness of machine learning algorithms. This iterative process often involves tasks such as handling missing values, removing duplicates, standardizing features, and encoding categorical variables. As noted by J. Brown et al. (2020) in their study on "Effective Data Preprocessing Techniques for Machine Learning," proper data reprocessing not only ensures data integrity but also contributes significantly to the success of machine learning endeavors, allowing models to generalize better and make more reliable predictions. Out of the total 303 instances in the dataset, a total of six (6) instances were identified to have missing values, representing approximately 2% of the entire dataset. Specifically, four (4) missing values were observed in the 'ca' attribute, and two (2) were observed in the 'thal' attribute.

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

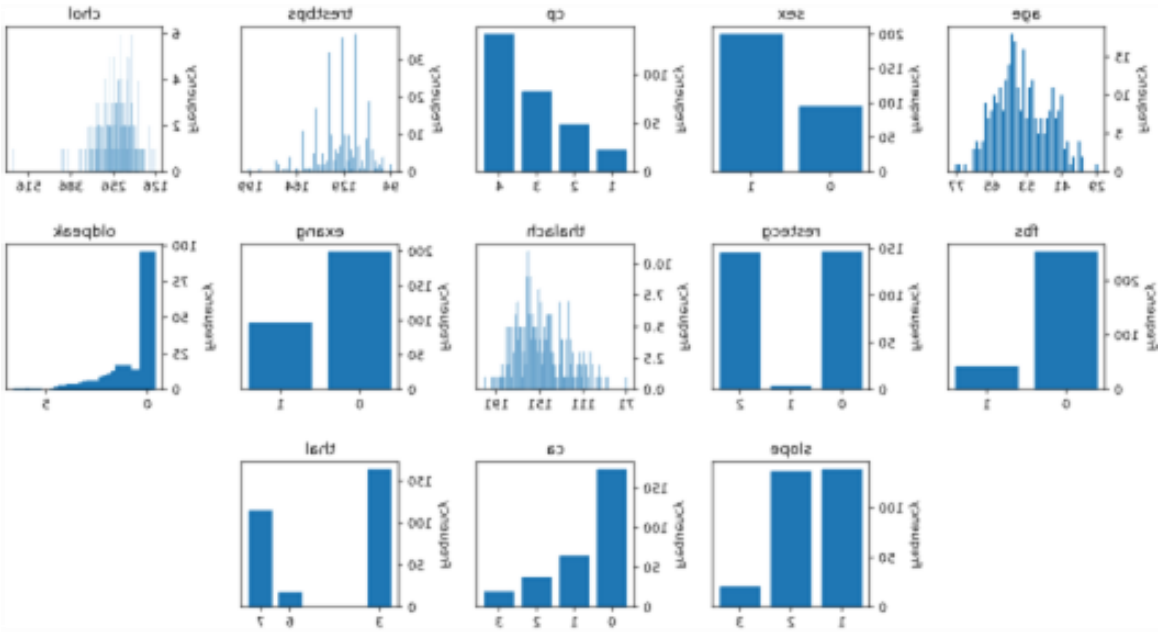

Figure 1. Histogram of Attributes.

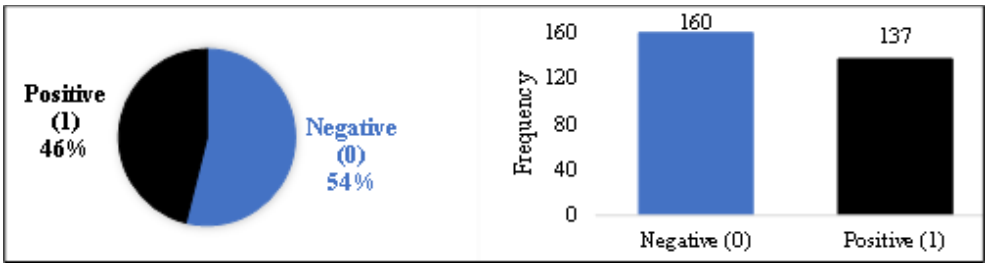

Figure 2. Target Class View.

11. LITERATURE ANALYSIS

Within this section, we have conducted an in-depth analysis and discussion of modern machine learning models created to predict heart disease. This analysis encompasses the years from 2019 to 2023. Our primary focus has been on evaluating the accuracy achieved by each method and the utilization of feature selection algorithms, as outlined in table 2. Furthermore, we have considered methods used in splitting data into training and testing sets, as well as the quantity of records utilized in the study. Table 2 clearly demonstrates that various data splitting techniques were employed in the literature. These methods encompass k-fold cross-validation and fixed splits, among others such as 73/27, 80/20, and 70/30. The choice of dataset split method depends on factors such as dataset size, data characteristics, and the goals of the machine learning experiment. It is fundamental to select a method that suits the specific requirements of your project to obtain meaningful and reliable results. Table 2 provides a summary of several research papers focused on heart disease prediction using machine learning methods. Each entry includes details about the author(s), publication year, employed methods or classifiers, evaluation parameters, the train/test split ratio, the number of records used, feature selection methods (if any), and the highest accuracy achieved. These studies collectively highlight the diverse range of machine learning algorithms, evaluation metrics,

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

and feature selection techniques employed in the field of heart disease prediction, with each study achieving notable levels of accuracy.

It can be observed from the finding that applying Relief, mRMR, Lasso, LLBFS feature selection models improves the performance of ML algorithms and demonstrated outstanding performance with perfect accuracy rate SVM 92.37 [50] , ET 94.41[53] . In contrast, using ML algorithms without feature selection models yielded the lowest accuracy rate K-NN 87 [52], and LR 86 [50] as shown in Table 2. Surprisingly, utilizing dimensionality reduction algorithms such as Chi-square, CHI-PCA, excelled in achieving optimal accuracy rate RF 98.7 [52]. Significant accuracy improvement has been observed after hyperparameter tuning of the classifiers SVM, it achieved an accuracy of 96.72% [56].

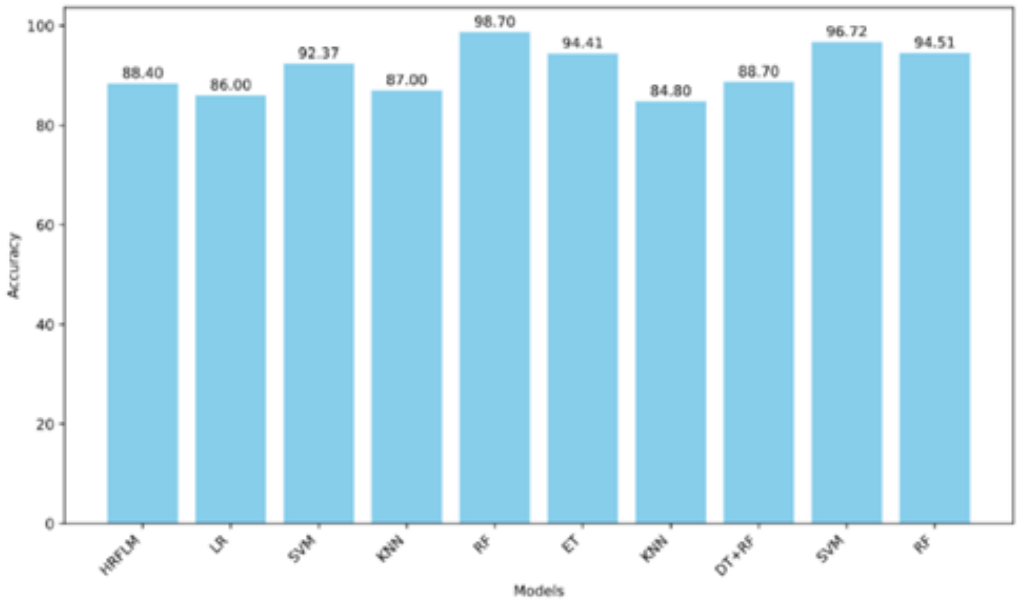

Figure 3. Accuracy of Modules.

Table 2. Comparative Result of Classification Techniques using Cleveland Heart Disease Data UCI.

Author	Year	Methods/ Classifiers	Train/Test	Highest Accuracy (%)
(Mohan et al., 2019) [48]	2019	NB, GLM, LR, DL, DT, RF, GBT, SVM, HRFLM(RF+LM)	70/30	HRFLM 88.4%
(Kannan & Vasanthi, 2019) [49]	2019	LR, RF, SVM, SGB	80/20	LR 86
(Li et al., 2020) [50	2020	SVM, LR, ANN, KNN NB, DT	70/30	SVM 92.37
(Archana Singh, 2020) [51]	2020	SVM, DT LR, K-NN	73/37	K-NN 87
(Gárate-Escamila et al., 2020) [52]	2020	DT, GBT, LOG, MPC, NB, RF	80/20	RF 98.7
(Muhammad et al., 2020) [53]	2020	KNN, DT, RF, NB, SVM, AB, ET, GB, LR, ANN	10-fold cross validation	ET 94.41
(Bharti et al., 2021) [54]	2021	LR, K-NN, SVM, RF DT	70/30	K-NN 84.8

International Journal of Research and Applied Technology

3(2) (2023) 267-282

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Author	Year	Methods/ Classifiers	Train/Test	Highest Accuracy (%)
(Kavitha et al., 2021) [55]	2021	DT, RF, Hybrid (DT+ R F)	70/30	Hybrid (DT+ R F) 88.7
(Saboor et al., 2022) [56]	2022	AB, LR, ET MNB,CART SVM, LDA	10-fold cross validation	SVM +hyper parameter 96.72
(Biswas et al., 2023) [57]	2023	RF, XGB LR, SVM, K-NN, RF, NB, DT	70/30	RF 94.51

12. LIMITATION AND FUTURE WORK

Although the ML models that were examined in the survey have shown encouraging outcomes, it's essential to acknowledge their significant shortcomings. Firstly, it is crucial to highlight that all of these models underwent training and assessment using comparatively small datasets. This particular aspect gives rise to concerns about the models' ability to deliver effective performance in practical, real-world situations where data tends to be extensive and diverse.

In addition to the previously mentioned limitations, it is indispensable to recognize another potential weakness in these surveyed ML models: their dependence on test data for hyperparameter selection. This practice poses a significant concern as it can lead to overfitting, a scenario where the model becomes excessively tailored to the specific test dataset used during hyperparameter tuning. Overfitting occurs when the model's hyperparameters are fine-tuned to maximize performance on the test data, but this might not necessarily result in improved generalization on unseen data.

In practical terms, it is advisable to partition the data into training, validation, and test sets. Ideally, hyperparameter tuning should be conducted on the validation set, ensuring that any enhancements in the model's performance are legitimate and not simply a result of fitting noise in the test data. Addressing this issue is crucial to guarantee that the ML models exhibit robustness and the ability to perform effectively across a broader spectrum of data, extending beyond the specific test dataset. This consideration underscores the significance of adhering to rigorous and principled practices in the development and evaluation of ML models.

Moreover, there's an aspect frequently disregarded in these investigations, which is the normalization of input data before training the models. Normalization stands as a pivotal preprocessing step in machine learning, as it ensures that input features are appropriately scaled to facilitate effective model training. The absence of information regarding the normalization techniques used raises concerns about the consistency and comparability of the outcomes outlined in Table 2. Various normalization methods, such as min-max scaling or z-score standardization, can yield a substantial influence on the performance of machine learning models.

The exclusion of these details from the study descriptions can be considered a limitation when evaluating the overall effectiveness of these models in predicting heart disease.

Improving the accuracy of machine learning algorithms in predicting heart disease is an ongoing challenge with continuous research efforts. An innovative way to potentially enhance accuracy that may not have been widely explored yet is Incorporate Genetic and Multi-Omics Data. While some studies have started to incorporate genetic data into heart disease prediction models, there is room for more comprehensive integration. Consider incorporating not only genetic information but also other omics data, such as transcriptomics, proteomics, and metabolomics. Utilize techniques from bioinformatics and genomics to identify genetic markers, expression patterns, and metabolic profiles associated with heart disease. Integrating

multi-omics data can provide a holistic view of an individual's health and susceptibility to heart disease.

13. CONCLUSION

This paper provides a comprehensive overview of multiple research studies focusing on the prediction and diagnosis of heart disease using various machine learning techniques. The comparison table summarizes key details from these studies, allowing for a quick comparison of their methodologies and results. Studies have diverse some area such as Diverse Range of Machine Learning Approaches, Evaluation Metrics, Feature Selection, Train/Test Split and Cross-Validation, High Accuracy Achieved.

Among the ML models surveyed, the highest accuracy rates were achieved using a combination of Chi-square with PCA (CHI-PCA) and Random Forest (RF), which achieved an impressive accuracy of 98.7% as shown in figure 3. Additionally, Support Vector Machine (SVM) achieved the highest accuracy of 96.72% after hyperparameter tuning. The omission of data normalization methods in certain research studies raises concerns regarding the reliability and uniformity of the outcomes reported. Also the potential for overfitting during hyperparameter tuning continue to be notable concerns within these studies. In future, focusing research efforts on Incorporate Genetic and Multi-Omics Data is considered an innovative way to potentially enhance accuracy of machine learning algorithms in predicting heart disease.

In conclusion, the combination of diverse machine learning techniques, rigorous evaluation metrics, and feature selection methods showcased in these studies demonstrates the significant potential of AI in enhancing heart disease diagnosis and prediction. Further research and collaboration between data scientists, clinicians, and healthcare institutions are crucial to advancing the field and translating these findings into practical clinical tools for improved patient care.

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed: September 7, 2023].

[2] American Heart Association, "Heart Disease," [Online]. Available: https://www.heart.org/en/health-topics/heart-disease. [Accessed: September 7, 2023].

[3] Yancy, C. W., Jessup, M., Bozkurt, B., Butler, J., Casey, D. E., Drazner, M. H., ... & Wilkoff, B. L. (2017). 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. *Circulation*, 136(6), e137-e161.

[4] Saeed, J. N., Abdulazeez, A. M., & Ibrahim, D. A., (2023). Automatic Facial Aesthetic Prediction Based on Deep Learning with Loss Ensembles. *Applied Sciences*, 13(17), p.9728.

[5] Saeed, J. N., Abdulazeez, A. M. & Ibrahim, D.A., 2023. An Ensemble DCNNs-Based Regression Model for Automatic Facial Beauty Prediction and Analyzation. *Traitement du Signal*, 40(1), p.55.

[6] Saeed, J. N., Abdulazeez, A. M., & Ibrahim, D. A. (2022, September). 2D Facial Images Attractiveness Assessment Based on Transfer Learning of Deep Convolutional Neural Networks. In *2022 4th International Conference on Advanced Science and Engineering (ICOASE)* (pp. 13-18). IEEE.

[7] Saeed, J. N., Abdulazeez, A. M., & Ibrahim, D. A. (2022, September). FIAC-Net: Facial image attractiveness classification based on light deep convolutional neural network. In *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-6). IEEE.

[8] Sadeeq, H.T., Ameen, S.Y. and Abdulazeez, A.M., 2022, November. Cancer Diagnosis based on Artificial Intelligence, Machine Learning, and Deep Learning. In *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)* (pp. 656-661). IEEE.

[9] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44-56.

[10] Rajkomar, A., Dean, J., & Kohane, I. (2018). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358.

[11] Smith, J. & Johnson, A. (2022). Machine learning algorithms in healthcare: Unveiling latent associations and prognostic insights. *Journal of Medical Informatics*, *45*(2), 123-136.

[12] UCI Machine Learning Repository. (Year). "Cleveland Heart Disease dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease. [Accessed: September 7, 2023].

[13] Alpaydin E., 2010. Introduction to Machine Learning, The MIT Press.

[14] Bishop C. M., 2006. Pattern Recognition and Machine Learning, Springer.

[15] T. Hastie, R. Tibshirani, & J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.

[16] Liu, B., Liu, C., Zhou, Y., Wang, D., & Dun, Y. (2023). An unsupervised chatter detection method based on AE and merging GMM and K-means. *Mechanical Systems and Signal Processing*, *186*, 109861.

[17] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[18] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[19] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

[20] Moore, S. K., Schneider, D., & Strickland, E. (2021). How Deep Learning Works: Inside the Neural Networks that Power Today's AI. *IEEE Spectrum*, *58*(10), 32-33.

[21] Zhang, H., Hao, K., Gao, L., Wei, B., & Tang, X. (2022). Optimizing deep neural networks through neuroevolution with stochastic gradient descent. *IEEE Transactions on Cognitive and Developmental Systems*, *15*(1), 111-121.

[22] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE access*, *7*, 53040-53065.

[23] Agrawal, P., Abutarboush, H. F., Ganesh, T., & Mohamed, A. W. (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019). *IEEE Access*, *9*, 26766-26791.

[24] Sadeeq, H. T., & Abdulazeez, A. M. (2023). Metaheuristics: A Review of Algorithms. International Journal of Online & Biomedical Engineering, 19(9).

[25] Sadeeq, H. T., & Abdulazeez, A. M. (2023, September). Car side impact design optimization problem using giant trevally optimizer. In Structures (Vol. 55, pp. 39-45). Elsevier.

[26] Sadeeq, H. T., & Abdulazeez, A. M. (2022). Giant trevally optimizer (GTO): A novel metaheuristic algorithm for global optimization and challenging engineering problems. *IEEE Access*, *10*, 121615-121640.

[27] Sadeeq, H. T., & Abdulazeez, A. M. (2022, September). Improved Northern Goshawk Optimization Algorithm for Global Optimization. In *2022 4th International Conference on Advanced Science and Engineering (ICOASE)* (pp. 89-94). IEEE.

[28] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

[29] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

[30] Ullah, F., Srivastava, G., Xiao, H., Ullah, S., Lin, J. C. W., & Zhao, Y. (2023). A Scalable Federated Learning Approach for Collaborative Smart Healthcare Systems with Intermittent Clients using Medical Imaging. IEEE Journal of Biomedical and Health Informatics.

[31] Yu, X., Qin, W., Lin, X., Shan, Z., Huang, L., Shao, Q., ... & Chen, M. (2023). Synergizing the enhanced RIME with fuzzy K-nearest neighbor for diagnose of pulmonary hypertension. *Computers in Biology and Medicine*, *165*, 107408.

[32] Syeds, S., & Thirupathy, P. (2023, May). A novel approach for precision and recall estimation for star rating online customers based on positive movie reviews using naive bayes algorithm over K-nearest neighbour algorithm. In *AIP Conference Proceedings* (Vol. 2655, No. 1). AIP Publishing.

[33] Akter, S., Siam, A. E., Monir, K. M., Mehedi, M. H. K., & Rasel, A. A. (2023, May). Bengali Movie Recommendation System using K Nearest Neighbor and Cosine Similarity. In *Proceedings of the 2023 9th International Conference on Computer Technology Applications* (pp. 25-29).

[34] John, G. H., & Langley, P. (2013). Estimating continuous distributions in Bayesian classifiers. *arXiv preprint arXiv:1302.4964*.

[35] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Xml retrieval. *Introduction to Information Retrieval*.

[36] Chebil, W., Wedyan, M., Alazab, M., Alturki, R., & Elshaweesh, O. (2023). Improving Semantic Information Retrieval Using Multinomial Naive Bayes Classifier and Bayesian Networks. *Information*, *14*(5), 272.

[37] Yati, J. D., Pamungkas, E. W., Kom, S., & Kom, M. Hate Speech Detection On Social Media Content In Javanese Language With Naive Bayes Algorithm (Doctoral dissertation, Universitas Muhammadiyah Surakarta).

[38] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees Belmont. *CA: Wadsworth International Group*.

[39] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81-106.

[40] Khalili, N., & Rastegar, M. A. (2023). Optimal cost-sensitive credit scoring using a new hybrid performance metric. *Expert Systems with Applications*, *213*, 119232.

[41] Qian, H., Ma, P., Gao, S., & Song, Y. (2023). Soft reordering one-dimensional convolutional neural network for credit scoring. *Knowledge-Based Systems*, *266*, 110414.

[42] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[43] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

[44] Adugna, T., Xu, W., & Fan, J. (2022). Comparison of random forest and support vector machine classifiers for regional land cover mapping using coarse resolution FY-3C images. Remote Sensing, 14(3), 574

[45] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

[46] Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

[47] Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(2), 458-475.

[48] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7, 81542-81554.

[49] Muppalaneni, N. B., Ma, M., Gurumoorthy, S., Kannan, R., & Vasanthi, V. (2019). Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. *Soft computing and medical bioinformatics*, 63-72.

[50] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE access*, *8*, 107562-107582.

[51] Rout, R. K. (2023, May). A Novel Grid Ann for Prediction of Heart Disease. In *2023 11th International Symposium on Electronic Systems Devices and Computing (ESDC)* (Vol. 1, pp. 1-6). IEEE.

[52] Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, 100330.

[53] Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Scientific reports*, *10*(1), 19747.

[54] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. *Computational intelligence and neuroscience*, *2021*.

[55] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In *2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333). IEEE.

[56] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems*, *2022*.

[57] Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., ... & Moni, M. A. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. *BioMed Research International*, *2023*.

[58] Zhou, H., Wang, X., & Zhu, R. (2022). Feature selection based on mutual information with correlation coefficient. *Applied Intelligence*, 1-18.

[59] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, *2*, 927312.

[60] Kako, N. A., & Abdulazeez, A. M. (2022). Peripapillary Atrophy Segmentation and Classification Methodologies for Glaucoma Image Detection: A Review. *Current Medical Imaging*, *18*(11), 1140-1159.