Journal of Advanced Health Informatics Research (JAHIR)
Vol. 1, No. 3, December 2023, pp. 157-174
ISSN: 2985-6124, DOI: 10.59247/jahir.v1i3.174

Exploration of Machine Learning Methods in Medical Disease Prediction: A Systematic Literature Review

Ria Suci Nurhalizah [1], Hadi Jayusman [2], Purwatiningsih [3]
[1,2] Program Studi Sistem Informasi, Universitas Harapan Bangsa, Purwokerto, Indonesia
[3] Program Studi Keperawatan Anestesiologi, Universitas Harapan Bangsa, Purwokerto, Indonesia

ARTICLE INFO

Article history:

Received November 28, 2023
Revised January 18, 2024
Published January 31, 2024

Keywords:

Machine Learning;
Medical;
Methods;
Disease;
Prediction

ABSTRACT

Exploration of Machine Learning methods in the systematic literature shows successful applications in disease diagnosis, disease prediction, and treatment planning. This literature only includes discussions on Classification methods consisting of Support Vector Machine(SVM), Naïve Bayes, Nearest Neighbors and Neural Network(NN) and Regression consisting of Decision Tree, Linear Regression, Random Forest Ensemble Methods, and Neural Network(NN). Clustering which consists of K-Means Clustering, Artificial Neural Network (ANN), Gaussian Mixture, Neural Network (NN) and Dimensionality reduction which consists of Principal Component Analysis (PCA). In the context of healthcare, the importance of sustainability, ethics, and data security are key factors. This research uses Systematic Literature Review (SLR) to explore Machine Learning methods in the medical context and recommends Support Vector Machine, Random Forest, and Neural Networks as effective methods. By exploring 300 papers and selecting 57 papers for discussion of machine learning methods in medical disease prediction. Method selection should be tailored to the dataset characteristics and disease prediction goals, while prioritizing ethics and data security in the application of Machine Learning in healthcare.

Corresponding Author:

Ria Suci Nuhalizah, Universitas Harapan Bangsa, Purwokerto, Indonesia
Email: riascnr02@gmail.com

1. INTRODUCTION

The use of Machine Learning in a medical context has become an increasing focus of research, along with its ability to process large and complex data to provide accurate predictions related to health diseases. Machine Learning is a subset of artificial intelligence (AI) in which tasks are performed by analyzing mathematical data, rather than being given specific programming instructions [1]. In other words, machine learning (ML) technology is a system designed to acquire knowledge autonomously without the need for direct instruction from the user. The capability of machine learning is its ability to understand patterns from data automatically without the need to explicitly specify how the data was generated. This helps us overcome these limitations and more effectively address the relationships between risk factors [2]. The importance of Machine Learning extends to many fields, utilizing its ability to learn from data and make predictions or decisions without the need for explicit programming. Machine Learning is proving invaluable in unearthing valuable insights, patterns, and relationships from a variety of complex datasets, transforming industries such as healthcare, finance, transportation, marketing, and other applications [3]. The success of machine learning technology is evident in a variety of sectors, playing an important role in image recognition, natural language processing, and anomaly detection [4].

In the healthcare sector, Machine Learning is widely used for diagnosing diseases, predicting disease forecasts, treatment planning, and customized treatment [5]. The importance of Information Technology (IT) in clinical risk management, decision support, and healthcare quality improvement is increasingly recognized [6]. In addition, applications of Machine Learning in the medical context have shown potential benefits, including cost reduction, time savings, improved quality of treatment, and better patient care [7]. The effectiveness of machine learning models in capturing complex and nonlinear relationships, surpassing the performance of traditional linear prediction models [8].

Medical disease risks, such as disease progression without obvious symptoms or individual risk factors, can be addressed by utilizing ML technology. This method enables complex data analysis, helping to detect patterns that are difficult to identify by traditional methods. Challenges in traditional diagnostics, such as limited diagnostic tools and lack of specialization, can be overcome with ML models that can process information quickly and provide more accurate predictions. By integrating ML into medical practice, it is expected to improve diagnostic efficiency and provide more effective solutions to medical disease risks.

In a recent study, the application of machine learning technology has successfully predicted many long-term diseases such as diabetes and performed well in several statistical measurements [9][10]. In addition, according to [11] The application of machine learning methods has shown promise in producing accurate predictions for the early stages of type 2 diabetes (T2DM). Prediction models using machine learning continue to be able to organize, classify, and correlate large amounts of multimodal data. This can provide support to clinicians in various diagnostic, prognostic, and therapeutic aspects, such as risk assessment, patient profiling, and resource allocation [12]. An important model is used for heart disease identification by referring to certain criteria and applying various methods from various studies to achieve optimal results. A study using a heart disease dataset achieved 81% accuracy through the application of the Random Forest algorithm. Another study utilizing a similar dataset and method, coupled with a Chi-Square feature selection process, achieved an accuracy rate of 83.7% [13]. Machine learning technology has also played a big role in creating tools for forecasting the risk of outbreaks, which are now often used in public health [14].

The advantages of using machine learning in systematic review can help reduce the tasks that must be done manually in the systematic review process [15]. Machine Learning involves the use of algorithms that can learn from complex patterns and relationships in data, instead of relying on rule-based approaches, thus allowing users to make more informed decisions [16]. Decision support systems that use machine learning can improve patient safety by better detecting errors, categorizing patients, and managing care or treatment. This will happen provided the system is used appropriately [17]. In addition, there are some weaknesses such as the system integrated in Machine Learning can be vulnerable to intrusion and manipulation through malicious attacks, where the attacker can take advantage of vulnerabilities in the system to erroneously classify harmless activities as malicious activities [18]. Unbalanced data challenges machine learning (ML) algorithms as they tend to correctly classify the majority class, but often misclassify the minority class [19].

The application of machine learning in healthcare faces several barriers that need to be taken seriously. One of the main obstacles is the lack of high-quality and complete data. Without adequate datasets, machine learning models tend to be less accurate and may produce unreliable predictions. In addition, it is important to consider the ethical aspects of using this technology. In the context of healthcare, there needs to be clarity on patient privacy and medical data security. The use of machine learning may also increase the risk of bias in decision-making, as models may reflect inequalities in the training data. Therefore, ethical measures such as transparency, accountability, and fairness are needed to ensure that the application of machine learning in healthcare provides maximum benefits without compromising ethical values. Exploration of machine learning methods in medical disease prediction should be done with caution and attention to these barriers and ethical aspects to ensure reliable results and sustainable application of the technology in healthcare.

2. METHODS

This research applies the Systematic Literature Review (SLR) method by collecting information from previous research using search keywords focused on Machine Learning Methods in the medical field. The literature search process was carried out through the Science Direct search engine within the last five years. The literature data that was successfully obtained was selectively chosen by researchers based on certain criteria. After finding suitable articles, the researcher analyzed and discussed based on points taken from the selection results, including research methodology, main findings, weaknesses, and potential for further development. This literature only includes discussions on Classification methods consisting of Support Vector Machine (SVM), Naïve Bayes, Nearest Neighbors and Neural Network (NN) and Regression consisting of Decision Tree, Linear Regression, Random Forest Ensemble Methods, and Neural Network (NN). Clustering which consists of K-Means Clustering, Artificial Neural Network (ANN), Gaussian Mixture, Neural Network (NN) and Dimensionality reduction which consists of Principal Component Analysis (PCA).

3. PREVIOUS LITERATURE

The application of Machine Learning in the prediction of various types of diseases with various methods has been described in a number of research papers. The following are some of the papers that discuss the topic:

Table 1. Previous Work

Paper Title	Author	Method	Year	Result	Advantages	Disadvantages
A support vector machine approach for identification of pleural effusion	Catur Edi Widodo [22]	Support Vector Machine (SVM)	2023	Identifying pleural effusion in thoracic images involves several stages, including region of interest (ROI) determination, segmentation, morphological operation, measurement of sharpness value and skew value, training, and testing. Testing was conducted with 50 thoracic images identified by doctors, and the results showed that the proposed method has an accuracy of 96%.	SVM has the advantage of being able to handle data that has many features, as it is capable of handling data with high dimensionality. In addition, SVM can also work well in cases where the decision boundary between data classes is non-linear.	The weakness of SVM lies in its limitation in handling very large datasets, as it requires a long computation time. In addition, SVM is also prone to overfitting if the C and gamma parameters are not set correctly.
A novel naive Bayes approach to identifying grooming behavior in the force-plate actometric platform	Collin J. Anderson [26]	Naive Bayes	2023	developed an automated paradigm to analyze grooming behavior in mice using a force-plate actometer. The naive Bayes approach used achieved 93.7% classification accuracy and an area under the receiver operating characteristic curve of 0.894.	Able to perform grooming analysis quickly, objectively, and automatically. Provides quantitative information on the force used in grooming. Can be used to perform grooming analysis in animal models of tick disorders and other psychiatric conditions.	Requires manual training to produce accurate classifications. Can only be used to analyze grooming behavior in animal models.
Analysis and prediction of	Almudena Sanjurjo-	Random Forest	2023	The authors analyzed 6030	provide valuable insights for	This study only analyzed a single

injury severity in single micromobility crashes with Random Forest	de-No [41]			single micromobility crashes that occurred in Spanish urban areas from 2016 to 2020. The authors used the Random Forest method to create a classification model with the aim of characterizing these accidents, predicting the severity of injuries, and identifying the main factors affecting them.	authorities in decision-making to improve micromobility safety and promote the creation of a more equitable and sustainable urban environment.	micromobility crash, so the results may not be used to analyze crashes involving other vehicles.
Application of ensemble models approach in anemia detection using images of the palpable palm	Peter Appiahene [44]	Ensemble methods	2023	Using palm images from 710 participants at a hospital in Ghana to develop a non-invasive machine learning model. The hybrid ensemble model, using techniques such as stacking, voting, boosting, and bagging, achieved an accuracy of 99.73%, indicating that the ensemble model is efficient for diagnosis or detection of medical diseases such as anemia.	This technique is non-invasive, so it does not require a large amount of money and time to detect anemia. It can be used in less developed communities, where medical resources and personnel are limited.	This technique can only be used to detect anemia. This technique requires a good quality palm image for accurate results.
Assessing the relationship between Body Mass Index and Bone Mineral Density in a clinical-based sample of Vietnamese	Trong Hung Nguyen [37]	linear regression	2024	examined the correlation between BMI and BMD in a generally healthy Vietnamese adult population. Data from 333 participants,	The use of generalized linear regression analysis methods can produce a predictive model that can be used to estimate BMD based on BMI. In addition, this study was	Firstly, this study was only conducted in a Vietnamese population aged 20-50 years, so the results cannot be generalized to other populations. Secondly, this study only

aged 20-50: A generalized linear regression analysis				who were aged 20-50 years and underwent nutritional assessment and BMD examination between January and November 2021. The study found that 7.7% of women and 4.6% of men had osteopenia of the spine, while 6.9% of women and 5.7% of men had osteopenia of the total hip.	conducted on a representative clinical sample of the Vietnamese population aged 20-50 years, so the results can be used to estimate BMD in a similar population.	evaluated the relationship between BMI and BMD, and did not consider other factors that may affect BMD such as physical activity, nutritional intake, and genetic factors.
Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making	Sadegh Ilbeigipour [46]	K-means Clustering	2022	explores the relationship between symptoms in patients who died from COVID-19 and patients who recovered. By filtering the data, the data was clustered using the K-means method and hierarchical clustering based on the Self Organizing Map (SOM) neural network. The results showed that patients who died from COVID-19 had high mean scores on various symptoms, although not all patients with these characteristics necessarily died. In addition, patient age is directly related to the duration of hospital stay, and elderly patients are	This method can help doctors and researchers make better medical decisions based on the geographic and demographic characteristics of patients. This method can help in understanding the patterns of COVID-19 spread around the world.	This method requires accurate and complete data to produce accurate results. This method requires significant time and resources to categorize COVID-19 cases. This method may not produce accurate results if the data used is incomplete or inaccurate.

				more likely to be placed in the intensive care unit (ICU).		
Disease progression modelling of Alzheimer's disease using probabilistic principal components analysis	Martin Saint-Jalmes [57]	Principal Components Analysis	2023	describe the temporal dynamics of biomarkers relevant to Alzheimer's disease using Principal Component Analysis (PCA). Principal Component Analysis helps identify the main components of biomarkers relevant to Alzheimer's and organize them into a new variable that reflects disease progression. The application of Principal Component Analysis helps understand key patterns in the data and provides a score that can be used as a pseudo-temporal indicator of disease progression.	The advantage of this model is its ability to combine multiple biomarkers into a single score that represents the progression of Alzheimer's disease in a patient. In addition, the model can also generate individual progression scores that can be updated over time and can be used as a pseudo-temporal scale to estimate biomarker evolution in Alzheimer's-affected patients.	Due to the heterogeneity of Alzheimer's disease, patients of the same age may experience different Alzheimer's-related changes. Therefore, disease progression models (DPMs) based on longitudinal mixed effects face challenges in the estimation of patient realigning time shifts. These time shifts are critical for meaningful biomarker modeling, but may affect the adjustment time or vary with missing data in co-estimated models.
Estimating the Health and Economic Outcomes of the Prevention of Mother-to-child Transmission of HIV Using a Decision Tree Model	QU Shui Ling [35]	Decision Tree	2019	A Decision Tree model was created to project the health and economic consequences on pregnant women of using PMTCT Option B+, focusing mainly on the impact of losses caused by abortion.	provides useful information on the health and economic impacts of preventing mother-to-child transmission of HIV. In this study, the authors show that prevention of mother-to-child HIV transmission can generate significant health and economic benefits. In addition, the authors also show that prevention of	only used a decision tree model to estimate the health and economic outcomes of preventing mother-to-child transmission of HIV. Therefore, the results of this study may not be fully accurate and reliable. In addition, this study only considered the health and economic impacts of preventing

					mother-to-child HIV transmission can reduce the number of new infections in children and can generate significant economic benefits.	mother-to-child HIV transmission and did not consider the social and psychological impacts of preventing mother-to-child HIV transmission.
Improving the impact of public health service delivery and research: a decision tree to aid evidence-based public health practice and research	Luke Wolfenden [34]	decision tree	2020	The use of decision trees created to provide support in the nursing field has been published, but few have addressed the aspect of applying research evidence to improve the effectiveness of public health programs and services.	provide practical guidance for health practitioners in implementing evidence-based health practices.	It does not provide concrete case examples and does not consider social and cultural factors that may influence the implementation of evidence-based health practices.
An ensemble nearest neighbor boosting technique for prediction of Parkinson's disease	K Aditya Shastry [30]	nearest neighbor	2023	detects PD at an early stage using ensemble techniques on the Parkinson's speech (PSV) dataset, which contains various voice recordings with 27 input features and one target feature combining k-Nearest Neighbor (k-NN) and Gradient Boosting (GB). The findings suggest that the designed ensemble method could be a promising approach for PD detection at an early stage.	Its high accuracy in predicting Parkinson's disease	Its high complexity and long computation time.
Edge Detection Algorithm of MRI Medical Image Based on	Shao Yunhong [50]	Artificial Neural Network	2022	To extract useful information from MRI medical images, effective image segmentation	uses the Canny edge detector as the training output and produces better image quality with about three	higher computational cost and subjective image quality can be further improved.

Artificial Neural Network				and proper use of edge operators are required. Using an edge detector based on the Canny Operator as the training output of an artificial neural network by training an artificial neural network. Results show the neural network-based edge detection method produces more complete MRI medical image edge information and processes it almost 3 times faster than traditional methods.	times faster processing time compared to other traditional algorithms such as Sobel and Canny edge detectors.	
Gaussian Mixture Model Implementation for Population Stratification Estimation from Genomics Data	Arif Budiarto [53]	Gaussian Mixture	2021	developed various methods to extract population stratification information from high-dimensional SNP (Single Nucleotide Polymorphisms) data. With the use of Principal Component Analysis (PCA) with Gaussian Mixture Model (GMM) as an unsupervised model to estimate population stratification from samples. Which is able to generate the probability distribution of each sample across the population, despite its	GMM can produce better results than other methods in estimating population stratification. GMM can generate probability distributions for each sample across the population. GMM can be used as an unsupervised model to estimate population stratification.	GMM requires longer computation time than other methods. GMM requires proper parameter selection to produce accurate results. GMM requires a large enough number of samples to produce accurate results.

				quality limitations.		
Neural networks applied to 12-lead electrocardiograms predict body mass index, visceral adiposity and concurrent cardiometabolic ill-health	Xinyang Li, PhD [32]	Neural networks	2021	developed a neural network (NN) model to predict body mass index (BMI) from ECG data and tested the hypothesis that differences between predicted and measured BMI may indicate fat or cardiometabolic disease. The NN model used 36,856 ECG data from the UK Biobank, with two architectures to estimate BMI. The model was tested separately for men and women (mean age 61±7 and 63±8 years; mean BMI 26±5 kg/m2 and 27±4 kg/m2). Results showed that the NN model could detect overweight/obesity with an average accuracy of 75% and 73% for males and females, respectively.	the use of neural network technology that can predict BMI with a fairly good level of accuracy.	This study used data from the UK Biobank which may not be representative of the wider population. In addition, this study only considered BMI and visceral adiposity in predicting cardiometabolic conditions, so other factors such as lifestyle and genetic factors were not considered.

4. RESULTS AND DISCUSSION

Machine learning methods are commonly used in medical disease prediction [20]:

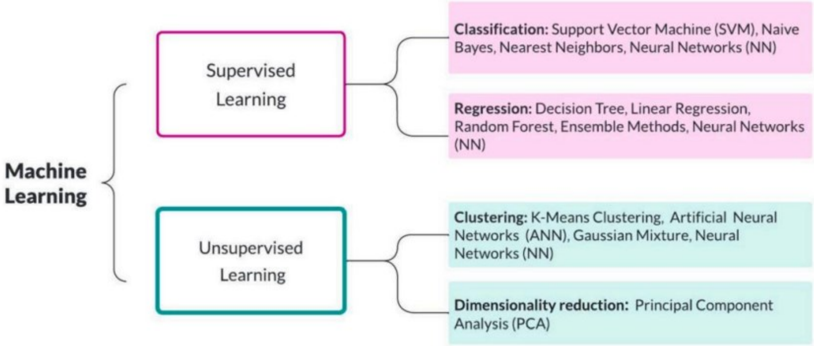
Machine Learning
Supervised Learning
Unsupervised Learning
Classification: Support Vector Machine (SVM), Naive Bayes, Nearest Neighbors, Neural Networks (NN)
Regression: Decision Tree, Linear Regression, Random Forest, Ensemble Methods, Neural Networks (NN)
Clustering: K-Means Clustering, Artificial Neural Networks (ANN), Gaussian Mixture, Neural Networks (NN)
Dimensionality reduction: Principal Component Analysis (PCA)

Fig. 1. Widely used machine learning algorithm

Basically, machine learning is divided into two, supervised learning and unsupervised learning. Supervised learning is a method in machine learning that utilizes data that has been labeled or datasets that are already known by the author. The labeled data serves to train the algorithm with guidance, so that the algorithm can classify or predict a case with a good level of accuracy. Supervised learning is usually used for two main things, namely Classification which consists of Support Vector Machine (SVM), Naïve Bayes, Nearest Neighbors and Neural Network (NN) and Regression which consists of Decision Tree, Linear Regression, Random Forest Ensemble Methods, and Neural Network (NN). Unsupervised learning is a method in machine learning that applies algorithms to analyze and find patterns in data without human intervention or assistance. In this context, no information is given about the expected results of an input to the algorithm, and the algorithm is tasked with finding patterns that may be present in the dataset. Generally, unsupervised learning is used for two main things: Clustering which consists of K-Means Clustering, Artificial Neural Network (ANN), Gaussian Mixture, Neural Network (NN) and Dimensionality reduction which consists of Principal Component Analysis (PCA).

4.1. Clasification

a. Support Vactor Mechine (SVM)

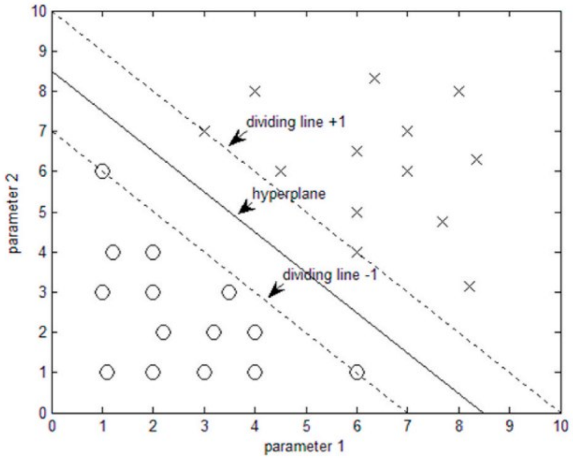

Fig. 2. Support Vector Machine

Support Vector Machine (SVM) is a powerful machine learning paradigm, used to perform supervised learning in the context of classification and regression analysis [21]. Support Vector Machine is an effective machine learning technique when used for data that is not included in the training set [22]. This algorithm works by finding the hyperplane that best separates data points into different classes, with the intention of maximizing the margin between classes [23]. They can help separate two classes or predict values based on training data.

b. Naive Bayes

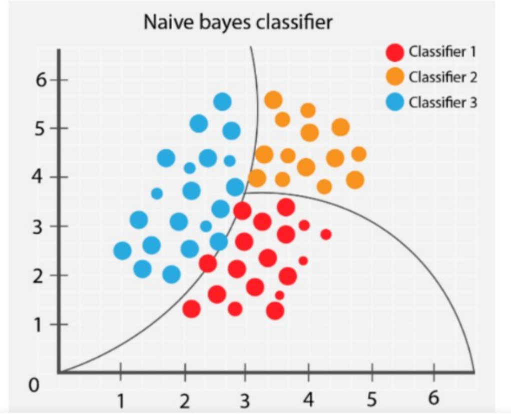

Fig. 3. Naïve Bayes

Naive Bayes is a way to categorize or classify things based on the basic idea of Bayes' Theorem. The basic idea is to use information about the probability of an event occurring based on previous information. So, Naive Bayes is used to predict categories or groups based on their probability [24]. In machine learning, Naive Bayes classification shows superior performance when compared to other traditional models that have been described in the literature [25].

c. Nearest Neighbors

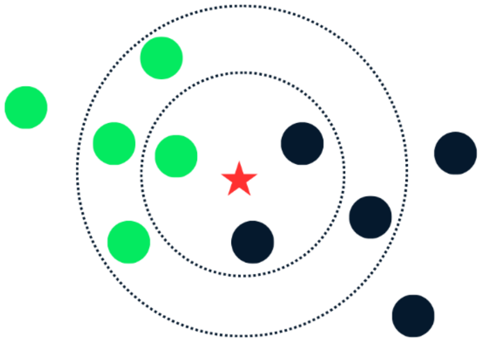

Fig. 3. K-Nearest Neighbors

Nearest Neighbors (k-NN) is a simple instance-based learning algorithm. It classifies or predicts the label of a new instance based on the class majority of its nearest neighbors in the feature space. KNN makes decisions locally by utilizing a number of nearby data objects, referred to as nearest neighbors or selectors, in the training data set [27]. The kNN classification has provided excellent results in categorizing medical information [28]. One of the most well-established supervised classification algorithms is the K-Nearest Neighbor (KNN) due to its ease and practicality in implementation [29]. The k-nearest neighbor method is a classification approach that does not require additional parameters. The method involves using the k nearest training samples from the dataset as input [30].

d. Neural Networks (NN)

Neural Networks are computational models inspired by the structure and function of the human brain. They consist of layers of interconnected nodes (neurons) and are used for various machine learning tasks. Neural Networks (NN) are increasingly pivotal in condition monitoring as they have the potential to enhance efficiency, diagnostic accuracy, and can be applied on a large scale [31].

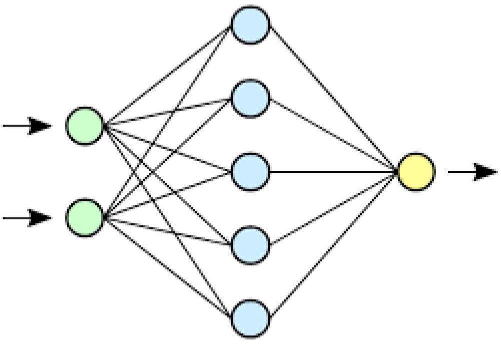

Fig. 4. Neural Network

4.2. Regresion

a. Decision Trees

In its history, Decision Tree (DT) has been recognized as an intelligent and highly useful technique in various areas such as Machine Learning, image processing, and pattern recognition. The Decision Tree model compares numbers with specific values during each testing step [6]. The Decision Tree approach creates rules that divide the subjects of analysis into several small groups, rather than classifying subjects randomly based on the researcher's decisions [33].

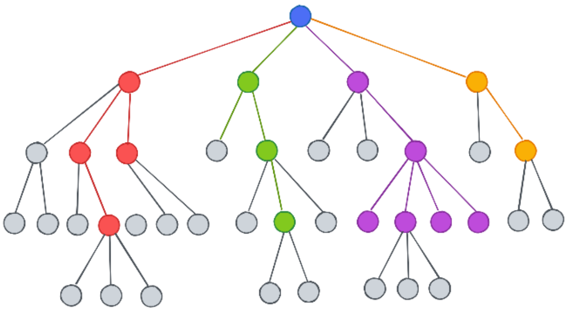

Fig. 5. Decision Tree

b. Linear Regression

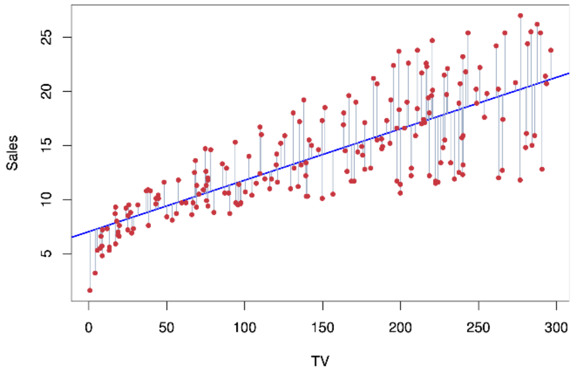

Fig. 6. Linier Regression

The statistical method commonly used to model cross-sectional data is linear regression analysis [36]. In regression modeling, there are two types of variables: the response variable (a variable that is influenced or its value depends on another variable) and the predictor variable (a variable believed to influence the response variable). In linear regression analysis, there are three objectives:

1. Formulating a regression model to understand the relationship between the dependent variable and the independent variable,
2. Conducting tests to determine whether there is an influence of the independent variable on the dependent variable, and
3. Making predictions of the dependent variable's values based on specified values of the independent variable.

c. Random Forest

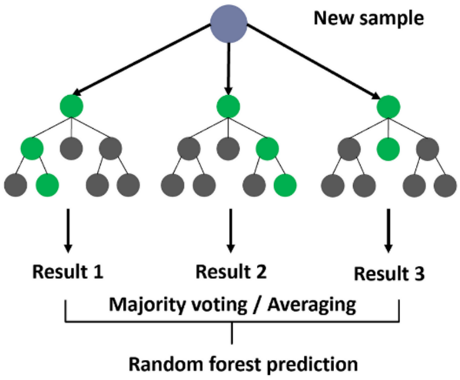

Fig. 7. Random Forest

Random Forest (RF) is a non-parametric machine learning technique used for classification and regression analysis. It consists of a collection of unpruned classification or regression trees built from randomly sampled training data [38]. Random Forest is one of the well-known machine learning algorithms that can be applied to various types of problems. This algorithm is characterized by its flexibility and ease of use [39]. Random Forest (RF) is one of the most commonly applied machine learning algorithms in the field of medicine and healthcare [40].

d. Ensemble Methods

The primary goal of using ensemble models is to generate predictions that are more robust than those produced by individual models. [42]. The combination of heterogeneous ensemble approaches, which integrates multiple machine learning algorithms, contributes to the development of more robust and accurate predictive models [43].

4.3. Clustering

a. K-means Clustering

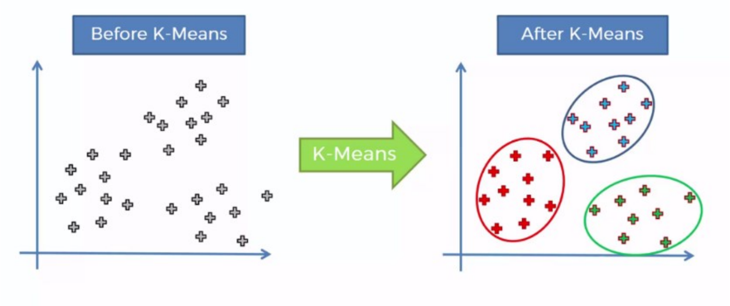

Fig. 8. K-means Clustering

K-means Clustering is an unsupervised machine learning algorithm used to partition data into k clusters based on similarity. It minimizes the sum of squared distances between data points and the assigned cluster centroids. The K-means algorithm involves an optimization process to ensure that samples within the same cluster exhibit a high level of similarity [45]. K-means clustering measures the distance between data objects and the cluster centroids, and data objects close to the centroid are assigned to the same category.

b. Artificial Neural Network (ANN)

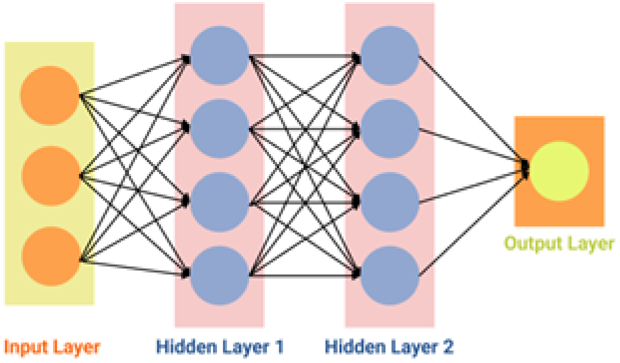

Fig. 9. Artificial Neural Network

Artificial Neural Network (ANN) is a machine learning algorithm inspired by the structure and function of the human brain [47]. Artificial Neural Network (ANN) is also frequently utilized in the medical field. Research is conducted on a number of individuals to create and ensure the effectiveness of the Artificial Neural Network algorithm. This algorithm aims to predict the likelihood of someone being hospitalized or experiencing death due to heart failure after a heart attack [48]. Artificial Neural Network has three layers, namely the input layer, hidden layer, and output layer. The input layer serves as the starting point to transfer data to the intermediate layer of neurons [49]. The intermediate layer is often referred to as the hidden layer, and the classic model of artificial neural networks consists of multiple hidden layers. The data is then transferred from the hidden layer of neurons to the output layer of neurons. The computational process runs through each layer via backpropagation, which is used to understand the complex relationship between the input and output layers.

c. Gaussian Mixture

The Gaussian Mixture Model is the sum of weights from Gaussian distributions, comprising mean and covariance values. The Gaussian Mixture Model (GMM) is a probability model that represents a mixture of Gaussian distributions. It is often used for clustering, where each Gaussian component represents a cluster. A one-dimensional Gaussian Mixture Model can be applied to describe time series data [51]. The purpose of using Gaussian Mixture Model is to predict and classify data into different classes based on probability distributions [52].

4.4. Dimensionality Reduction

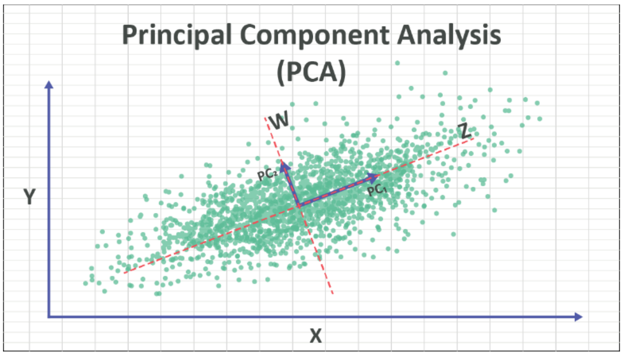

Fig. 10. Principal Component Analysis

Principal Component Analysis is one of the techniques commonly used in the literature to reduce the complexity of objects [54]. It is a statistical tool used to analyze complex data, or more specifically, to identify redundant variables that do not provide additional information but rather make the data more intricate. Principal Component Analysis involves transforming several variables into a set of more comprehensive variables by applying the concept of dimensionality reduction [55]. The goal of Principal Component Analysis is to calculate the spectral shape that best approximates, in the least squared sense, the entire spectrum present in a dataset [56].

5. CONCLUSION

In the discussion above, it can be concluded that in the exploration of Machine Learning methods for predicting medical diseases, several methods have been widely used and successful in various studies. Some popular classification methods include Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (KNN), and Neural Networks based on research consisting of 57 papers. On the other hand, regression methods such as Decision Trees, Linear Regression, and Random Forest also show effectiveness in disease prediction and regression analysis. The implementation of ensemble methods such as stacking, bagging, voting, and boosting proves their strength in improving overall prediction accuracy and performance. In terms of clustering, K-means Clustering and Artificial Neural Network (ANN) have proven to be effective, as well as Gaussian Mixture Model (GMM). To address data complexity, Principal Component Analysis (PCA) as a dimensionality reduction method has been widely used.

6. SUMMARY POINTS

6.1. The Use of Machine Learning in Health: Machine Learning is gaining increasing attention in disease prediction due to its ability to process large and complex data, providing accurate health-related predictions.

6.2. Definition of Machine Learning: Machine Learning is a part of artificial intelligence where machines can learn through mathematical analysis of data without specific programming instructions.

6.3. Importance of Machine Learning: Machine Learning plays a crucial role in various sectors, including health, with its ability to discover patterns and relationships from complex datasets.

6.4. Applications of Machine Learning in Health: Machine Learning is used for disease diagnosis, predicting disease outcomes, treatment planning, and personalized medicine.

6.5. Success of Machine Learning Technology: Machine Learning is effective in identifying images, natural language processing, anomaly detection, and offers potential benefits such as cost reduction and improved quality of treatment.

6.6. Limitations of Machine Learning Usage: Main obstacles include limited availability of high-quality data, the risk of bias in decision-making, and ethical concerns related to patient privacy.

6.7. Exploration of Machine Learning Methods: In literature, classification methods such as SVM, Naive Bayes, KNN, and Neural Networks have proven successful. Regression, ensemble methods, and clustering are also effective.

6.8. Importance of Systematic Literature Review: This research applies Systematic Literature Review (SLR) to gather information from previous studies, focusing on Machine Learning methods in the medical field.

6.9. Popular Methods: SVM, Random Forest, and Neural Networks emerge as popular methods. The application of ensemble methods and dimensionality reduction techniques like PCA is also significant.

6.10. Recommendations for Method Usage: The selection of methods should be tailored to the characteristics of the dataset and the goals of disease prediction. Ethics and data security should be prioritized in the implementation of Machine Learning in healthcare.

REFERENCES

[1] S. S. Rana, B. Nath, P. K. Chaudhari, and S. Vichare, “Cervical Vertebral Maturation Assessment using various Machine Learning techniques on Lateral cephalogram: A systematic literature review,” *Journal of Oral Biology and Craniofacial Research*, vol. 13, no. 5. Elsevier B.V., pp. 642–651, Sep. 01, 2023. doi: 10.1016/j.jobcr.2023.08.005.

[2] I. Ismail, P. J. A. Stam, F. R. M. Portrait, A. van Witteloostuijn, and X. Koolman, “Addressing unanticipated interactions in risk equalization: A machine learning approach to modeling medical expenditure risk,” *Econ Model*, vol. 130, Jan. 2024, doi: 10.1016/j.econmod.2023.106564.

[3] O. Nooruldeen, M. R. Baker, A. M. Aleesa, A. Ghareeb, and E. H. Shaker, "Strategies for predictive power: Machine learning models in city-scale load forecasting," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 6, Dec. 2023, doi: 10.1016/j.prime.2023.100392.

[4] Z. Sun, G. An, Y. Yang, and Y. Liu, "Optimized machine learning enabled intrusion detection 2 system for internet of medical things," *Franklin Open*, vol. 6, p. 100056, Mar. 2024, doi: 10.1016/j.fraope.2023.100056.

[5] L. Xiao *et al.*, "Predictive model for early death risk in pediatric hemophagocytic lymphohistiocytosis patients based on machine learning," *Heliyon*, vol. 9, no. 11, p. e22202, Nov. 2023, doi: 10.1016/j.heliyon.2023.e22202.

[6] M. M. Hossain, M. A. Kashem, N. M. Nayan, and M. A. Chowdhury, "A Medical Cyber-physical system for predicting maternal health in developing countries using machine learning," *Healthcare Analytics*, vol. 5, Jun. 2024, doi: 10.1016/j.health.2023.100285.

[7] J. Allgaier, L. Mulansky, R. L. Draelos, and R. Pryss, "How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare," *Artif Intell Med*, vol. 143, Sep. 2023, doi: 10.1016/j.artmed.2023.102616.

[8] C. Montorsi, A. Fusco, P. Van Kerm, and S. P. A. Bordas, "Predicting depression in old age: Combining life course data with machine learning," *Econ Hum Biol*, vol. 52, Jan. 2024, doi: 10.1016/j.ehb.2023.101331.

[9] S. S. Bhat, M. Banu, G. A. Ansari, and V. Selvam, "A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms," *Healthcare Analytics*, vol. 4. Elsevier Inc., Dec. 01, 2023. doi: 10.1016/j.health.2023.100273.

[10] S. Jangili, H. Vavilala, G. S. B. Boddeda, S. M. Upadhyayula, R. Adela, and S. R. Mutheneni, "Machine learning-driven early biomarker prediction for type 2 diabetes mellitus associated coronary artery diseases," *Clin Epidemiol Glob Health*, vol. 24, Nov. 2023, doi: 10.1016/j.cegh.2023.101433.

[11] G. S, V. S. Reddy, and M. R. Ahmed, "Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus," *Measurement: Sensors*, p. 100983, Dec. 2023, doi: 10.1016/j.measen.2023.100983.

[12] T. Frondelius *et al.*, "Early prediction of ventilator-associated pneumonia with machine learning models: A systematic review and meta-analysis of prediction model performance☆," *Eur J Intern Med*, 2023, doi: 10.1016/j.ejim.2023.11.009.

[13] P. S. Asih, Y. Azhar, G. W. Wicaksono, and D. R. Akbi, "Interpretable Machine Learning Model For Heart Disease Prediction," *Procedia Comput Sci*, vol. 227, pp. 439–445, 2023, doi: 10.1016/j.procs.2023.10.544.

[14] T. Zhang, F. Rabhi, X. Chen, H. Paik, and C. R. MacIntyre, "A machine learning-based universal outbreak risk prediction tool," *Comput Biol Med*, p. 107876, Dec. 2023, doi: 10.1016/j.compbiomed.2023.107876.

[15] B. G. Pijls, "Machine Learning assisted systematic reviewing in orthopaedics," *J Orthop*, vol. 48, pp. 103–106, Feb. 2024, doi: 10.1016/j.jor.2023.11.051.

[16] S. Jahandideh, G. Ozavci, B. W. Sahle, A. Z. Kouzani, F. Magrabi, and T. Bucknall, "Evaluation of machine learning-based models for prediction of clinical deterioration: A systematic literature review," *International Journal of Medical Informatics*, vol. 175. Elsevier Ireland Ltd, Jul. 01, 2023. doi: 10.1016/j.ijmedinf.2023.105084.

[17] J. W. Asare, P. Appiahene, and E. T. Donkoh, "Detection of anaemia using medical images: A comparative study of machine learning algorithms – A systematic literature review," *Informatics in Medicine Unlocked*, vol. 40. Elsevier Ltd, Jan. 01, 2023. doi: 10.1016/j.imu.2023.101283.

[18] O. Alshaikh, S. Parkinson, and S. Khan, "Exploring Perceptions of Decision-Makers and Specialists in Defensive Machine Learning Cybersecurity Applications: The Need for a Standardised Approach," *Comput Secur*, p. 103694, Dec. 2023, doi: 10.1016/j.cose.2023.103694.

[19] A. X. Wang, S. S. Chukova, and B. P. Nguyen, "Synthetic minority oversampling using edited displacement-based k-nearest neighbors," *Appl Soft Comput*, vol. 148, Nov. 2023, doi: 10.1016/j.asoc.2023.110895.

[20] G. Kantayeva, J. Lima, and A. I. Pereira, "Application of machine learning in dementia diagnosis: A systematic literature review," *Heliyon*, vol. 9, no. 11, Nov. 2023, doi: 10.1016/j.heliyon.2023.e21626.

[21] J. A. Warwicker and S. Rebennack, "Support vector machines within a bivariate mixed-integer linear programming framework," *Expert Syst Appl*, vol. 245, p. 122998, Jul. 2024, doi: 10.1016/j.eswa.2023.122998.

[22] C. E. Widodo, K. Adi, and R. Gernowo, "A support vector machine approach for identification of pleural effusion," *Heliyon*, p. e22778, Nov. 2023, doi: 10.1016/j.heliyon.2023.e22778.

[23] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, "A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms," *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2023.100185.

[24] M. Vishwakarma and N. Kesswani, "A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection," *Decision Analytics Journal*, vol. 7, Jun. 2023, doi: 10.1016/j.dajour.2023.100233.

[25] N. Deepa, J. Sathya Priya, and T. Devi, "Towards applying internet of things and machine learning for the risk prediction of COVID-19 in pandemic situation using Naive Bayes classifier for improving accuracy," *Mater Today Proc*, vol. 62, pp. 4795–4799, Jan. 2022, doi: 10.1016/j.matpr.2022.03.345.

[26] C. J. Anderson *et al.*, "A novel naïve Bayes approach to identifying grooming behaviors in the force-plate actometric platform," *J Neurosci Methods*, vol. 403, Mar. 2024, doi: 10.1016/j.jneumeth.2023.110026.

[27] S. Suyanto, P. E. Yunanto, T. Wahyuningrum, and S. Khomsah, “A multi-voter multi-commission nearest neighbor classifier,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6292–6302, Sep. 2022, doi: 10.1016/j.jksuci.2022.01.018.

[28] F. J. Gomez-Gil, V. Martínez-Martínez, R. Ruiz-Gonzalez, L. Martínez-Martínez, and J. Gomez-Gil, “Vibration-based monitoring of agro-industrial machinery using a k-Nearest Neighbors (kNN) classifier with a Harmony Search (HS) frequency selector algorithm,” *Comput Electron Agric*, vol. 217, p. 108556, Feb. 2024, doi: 10.1016/j.compag.2023.108556.

[29] N. Zamri *et al.*, “River quality classification using different distances in k-nearest neighbors algorithm,” in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 180–186. doi: 10.1016/j.procs.2022.08.022.

[30] K. A. Shastry, “An ensemble nearest neighbor boosting technique for prediction of Parkinson’s disease,” *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2023.100181.

[31] N. Herwig and P. Borghesani, “Explaining deep neural networks processing raw diagnostic signals,” *Mech Syst Signal Process*, vol. 200, Oct. 2023, doi: 10.1016/j.ymssp.2023.110584.

[32] X. Li, K. H. K. Patel, L. Sun, N. S. Peters, and F. S. Ng, “Neural networks applied to 12-lead electrocardiograms predict body mass index, visceral adiposity and concurrent cardiometabolic ill-health,” *Cardiovasc Digit Health J*, vol. 2, no. 6, pp. S1–S10, Dec. 2021, doi: 10.1016/j.cvdhj.2021.10.003.

[33] J. Oh and B. Kim, “Prediction Model for Demands of the Health Meteorological Information Using a Decision Tree Method,” 2010.

[34] L. Wolfenden *et al.*, “Improving the impact of public health service delivery and research: a decision tree to aid evidence-based public health practice and research,” *Australian and New Zealand Journal of Public Health*, vol. 44, no. 5. Wiley-Blackwell, pp. 331–332, Oct. 01, 2020. doi: 10.1111/1753-6405.13023.

[35] S. L. QU, A. L. WANG, X. P. PAN, Q. WANG, L. X. DOU, and T. ZHANG, “Estimating the Health and Economic Outcomes of the Prevention of Mother-to-child Transmission of HIV Using a Decision Tree Model,” *Biomedical and Environmental Sciences*, vol. 32, no. 1, pp. 68–74, Jan. 2019, doi: 10.3967/bes2019.011.

[36] S. D. Permai and H. Tanty, “Linear regression model using bayesian approach for energy performance of residential building,” in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 671–677. doi: 10.1016/j.procs.2018.08.219.

[37] T. H. Nguyen *et al.*, “Assessing the relationship between Body Mass Index and Bone Mineral Density in a clinical-based sample of Vietnamese aged 20–50: A generalized linear regression analysis,” *Human Nutrition & Metabolism*, vol. 35, p. 200241, Mar. 2024, doi: 10.1016/j.hnm.2024.200241.

[38] G. Grekousis, Z. Feng, I. Marakakis, Y. Lu, and R. Wang, “Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach,” *Health Place*, vol. 74, Mar. 2022, doi: 10.1016/j.healthplace.2022.102744.

[39] M. Mojsilović, R. Cvejić, S. Pepić, D. Karabašević, M. Saračević, and D. Stanujkić, “Statistical evaluation of the achievements of professional students by combination of the random forest algorithm and the ANFIS method,” *Heliyon*, vol. 9, no. 11, Nov. 2023, doi: 10.1016/j.heliyon.2023.e21768.

[40] J.-J. Chen, L.-F. Liu, S.-M. Chang, and C.-P. Lu, “Identifying the top determinants of psychological resilience among community older adults during COVID-19 in Taiwan: A random forest approach,” *Machine Learning with Applications*, vol. 14, p. 100494, Dec. 2023, doi: 10.1016/j.mlwa.2023.100494.

[41] A. Sanjurjo-de-No, A. M. Pérez-Zuriaga, and A. García, “Analysis and prediction of injury severity in single micromobility crashes with Random Forest,” *Heliyon*, vol. 9, no. 12, Dec. 2023, doi: 10.1016/j.heliyon.2023.e23062.

[42] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, “Leveraging Bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach,” *Heliyon*, p. e24188, Jan. 2024, doi: 10.1016/j.heliyon.2024.e24188.

[43] K. Sriprateep *et al.*, “Heterogeneous ensemble machine learning to predict the asiaticoside concentration in centella asiatica urban,” *Intelligent Systems with Applications*, vol. 21, p. 200319, Mar. 2024, doi: 10.1016/j.iswa.2023.200319.

[44] P. Appiahene *et al.*, “Application of ensemble models approach in anemia detection using images of the palpable palm,” *Med Nov Technol Devices*, vol. 20, Dec. 2023, doi: 10.1016/j.medntd.2023.100269.

[45] X. Wang, Z. Shao, Y. Shen, and Y. He, “Research on fast marking method for indicator diagram of pumping well based on K-means clustering,” *Heliyon*, vol. 9, no. 10, Oct. 2023, doi: 10.1016/j.heliyon.2023.e20468.

[46] S. Ilbeigipour, A. Albadvi, and E. Akhondzadeh Noughabi, “Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making,” *Inform Med Unlocked*, vol. 32, Jan. 2022, doi: 10.1016/j.imu.2022.101005.

[47] P. G, V. R. Chintala, T. Reddy, and R. T, “User-Cloud-based Ensemble Framework for Type-2 Diabetes Prediction with Diet Plan Suggestion,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, p. 100423, Jan. 2024, doi: 10.1016/j.prime.2024.100423.

[48] J. K. Chaw *et al.*, “A predictive analytics model using machine learning algorithms to estimate the risk of shock development among dengue patients,” *Healthcare Analytics*, vol. 5, Jun. 2024, doi: 10.1016/j.health.2023.100290.

[49] A. Shah *et al.*, “A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN),” *Clinical eHealth*, vol. 6. KeAi Communications Co., pp. 76–84, Dec. 01, 2023. doi: 10.1016/j.ceh.2023.08.002.

[50] S. Yunhong, Y. Shilei, Z. Xiaojing, and Y. Jinhua, "Edge Detection Algorithm of MRI Medical Image Based on Artificial Neural Network," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 136–144. doi: 10.1016/j.procs.2022.10.021.

[51] E. Külah, Y. M. Çetinkaya, A. G. Özer, and H. Alemdar, "COVID-19 forecasting using shifted Gaussian Mixture Model with similarity-based estimation," *Expert Syst Appl*, vol. 214, Mar. 2023, doi: 10.1016/j.eswa.2022.119034.

[52] M. Hamdi, I. Hilali-Jaghdam, B. E. Elnaim, and A. A. Elhag, "Forecasting and classification of new cases of COVID 19 before vaccination using decision trees and Gaussian mixture model," *Alexandria Engineering Journal*, vol. 62, pp. 327–333, Jan. 2023, doi: 10.1016/j.aej.2022.07.011.

[53] A. Budiarto, B. Mahesworo, A. A. Hidayat, I. Nurlaila, and B. Pardamean, "Gaussian Mixture Model Implementation for Population Stratification Estimation from Genomics Data," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 202–210. doi: 10.1016/j.procs.2021.12.026.

[54] H. Sharma, G. Mandil, É. Monnier, E. Cor, and P. Zwolinski, "Sizing a hybrid hydrogen production plant including life cycle assessment indicators by combining NSGA-III and principal component analysis (PCA)," *Energy Conversion and Management: X*, vol. 18, Apr. 2023, doi: 10.1016/j.ecmx.2023.100361.

[55] K. Zhang, Z. Chen, L. Yang, and Y. Liang, "Principal component analysis (PCA) based sparrow search algorithm (SSA) for optimal learning vector quantized (LVQ) neural network for mechanical fault diagnosis of high voltage circuit breakers," *Energy Reports*, vol. 9, pp. 954–962, Mar. 2023, doi: 10.1016/j.egyr.2022.11.118.

[56] N. Fairley, P. Bargiela, W. M. Huang, and J. Baltrusaitis, "Principal Component Analysis (PCA) unravels spectral components present in XPS spectra of complex oxide films on iron foil," *Applied Surface Science Advances*, vol. 17, Oct. 2023, doi: 10.1016/j.apsadv.2023.100447.

[57] M. Saint-Jalmes *et al.*, "Disease progression modelling of Alzheimer's disease using probabilistic principal components analysis," *Neuroimage*, vol. 278, Sep. 2023, doi: 10.1016/j.neuroimage.2023.120279.