A Survey of Classification Algorithms in Supervised Machine Learning

Mageshwari G.[1], Dr. Ramar K.[2], Monica R. Lakshmi[3]
Assistant Professor, R.M.K. College of Engineering and Technology
Professor, R.M.K. College of Engineering and Technology
Assistant Professor, R.M.D. Engineering College

Abstract—Machine learning is crucial in enhancing predictive and diagnostic capabilities across multiple sectors. Professionals can use it to identify potential conditions and assess the risks associated with different intervention strategies. Machine Learning methods have shown significant potential in enhancing disease detection by offering accurate, efficient, and automated diagnostic capabilities. Supervised machine learning is a widely used approach in artificial intelligence that enables systems to learn from labeled data and make accurate predictions. This paper explores various supervised learning techniques, including classification models, which are applied across diverse domains such as healthcare, finance, and natural language processing. This study focuses on the approaches and the applications of supervised learning and highlights its benefits, and discusses ongoing challenges and future directions for improving machine learning-based healthcare solutions.

Keywords: Health Care, Machine Learning, Supervised Learning

*I. I*NTRODUCTION

Artificial Intelligence is the ability to automatically learn and improve based on experience, which is machine learning. Complex tasks can be accomplished with Artificial Intelligence systems using the same approach humans take to solving them. Machine learning has a tremendous role everywhere. The Machine learning of AI uses techniques to learn more about the data, recognize patterns from data and apply them to make better decisions. The rapid advancement of machine learning has significantly transformed the healthcare industry, particularly in disease detection and diagnosis. Traditional diagnostic methods rely heavily on human expertise, which can be time-consuming and prone to errors. Supervised machine learning, a subset of artificial intelligence, addresses these challenges by utilizing labeled medical data to train predictive models. These models learn from past cases to identify patterns and make accurate disease classifications. Supervised learning techniques such as Logistic Regression, Support Vector Machines (SVM), Random Forests, and Deep Neural Networks (DNN) have shown great promise in medical applications, including the detection of diseases like cancer, diabetes, and cardiovascular conditions. These models are effective in classifying patients based on various features, such as medical test results, demographic data, or medical images. Logistic Regression is simple and interpretable but limited by its assumption of linear relationships. SVMs, on the other hand, are powerful for complex classifications but require careful parameter tuning and can be computationally expensive. Despite their advantages, these models face significant challenges, such as a lack of data, especially for rare diseases, and the need for proper feature selection to avoid overfitting. Hyperparameter tuning, which ensures optimal model performance, can be time-consuming and computationally expensive. Furthermore, model interpretability remains a concern, particularly in healthcare, where understanding why a model makes a certain prediction is crucial for trust. Ethical issues, such as ensuring patient privacy and securing sensitive medical data, also pose significant hurdles, particularly given the strict regulations governing healthcare data. Additionally, these models must generalize well across diverse populations, as training on a limited or biased dataset can lead to unfair or inaccurate predictions. Thus, while supervised learning holds tremendous potential in improving healthcare outcomes, addressing these challenges is key to its effective and ethical implementation. The rest of the Chapter is organized as follows: Chapter 2 reviews related work. Chapter 3 discusses the Challenges. Chapter 4 explains the approaches of classification techniques. Chapter 5 concludes the paper.

II. RELATED WORKS

The use of Artificial Neural Networks (ANN) combined with ECG and respiratory signals to predict bradycardia in neonates. The challenge lies in generalization issues due to rapid heart rate changes and limited input signals. Future research suggests exploring alternative ML models, incorporating more physiological signals, and handling clustered bradycardic episodes for improved accuracy [1]. AI-powered Clinical Decision Support Systems (CDSS) for cardiovascular disease risk assessment, diagnosis, treatment, and monitoring. Key challenges include data quality,

privacy, security, clinical validation, AI adoption, and ethical concerns. To enhance AI applications in cardiovascular care, researchers propose improving model accuracy, integrating AI with wearable devices, expanding applications, and conducting large-scale trials [2].

Ensemble techniques such as bagging, boosting, and stacking to predict coronary heart disease. Challenges include data quality, computational complexity, and lower recall scores. Future improvements include validation using clinical data, exploring deep learning models, and optimizing feature selection for better disease prediction accuracy [3]. Various ML models, including KNN, Decision Trees, Random Forest, SVM, and Logistic Regression, are used for heart disease prediction. However, the study highlights challenges such as data quality, computational complexity, overfitting, and feature selection. Future directions involve integrating real-time clinical data, exploring deep learning techniques, and incorporating wearable device data for continuous monitoring [4].

ML models such as Decision Trees, Random Forest, XGBoost, SVM, and MLP, trained on a combined dataset from multiple sources for cardiovascular disease diagnosis. Challenges include data quality issues, high computational cost, and limited generalizability. Future enhancements focus on developing explainable AI, validating models with real-world clinical data, integrating with wearable technology, and exploring deep learning techniques [5]. A hybrid model combining 1D CNN and LSTM with an output correction mechanism is proposed for neonatal bradycardia detection. The study faces challenges such as dataset limitations, feature selection, comparison with other ML models, and generalization issues. Future work suggests testing the model on diverse datasets, improving feature engineering, and optimizing the correction mechanism for better reliability [6]. ML models, including ANN, Logistic Regression, SVM, Random Forest, and Ensemble Voting, to predict heart disease. Major challenges include clinical integration, web accessibility, model expansion, and real-time monitoring. Future improvements involve combining Random Forest with AdaBoost, implementing IoT-based real-time monitoring, and expanding the model's usability for broader healthcare applications [7].

Healthcare Predictive Analytics investigates ML (Random Forest, Decision Trees, SVM, KNN) and deep learning (CNN, LSTM, RCNN) models for healthcare prediction. Key challenges include data privacy and security, explainability of AI models, computational complexity, and generalization. Future research aims to develop hybrid ML-DL models, improve real-time monitoring using IoT, and focus on rare disease detection [8]. ML techniques, including SVM, ANN, Decision Trees, CNN, and LSTM, applied to disease prediction, medical imaging, and decision support. Challenges include data privacy, model interpretability, dataset diversity, and healthcare integration. Future advancements focus on explainable AI, federated learning, AI-driven telemedicine, and rare disease prediction to improve healthcare AI applications [9].

1 CHALLENGES

Machine learning (ML) and artificial intelligence (AI) have been widely applied in cardiovascular and healthcare analytics for disease prediction, diagnosis, and monitoring. Studies have utilized various ML models, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, Random Forest, XGBoost, and ensemble techniques, to predict and detect conditions such as bradycardia, coronary heart disease, and general heart disease. Deep learning approaches like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models have also been explored, particularly in neonatal bradycardia detection and healthcare predictive analytics.

Challenges across these studies include data quality, computational complexity, overfitting, model generalization, and ethical concerns such as privacy and security. Future research directions focus on improving model accuracy, integrating AI with wearable devices, enhancing real-time monitoring, and employing explainable AI techniques to ensure reliability in clinical settings.

2 APPROACHES

The classification approaches discussed include five major algorithms used in data mining and machine learning. Decision Tree Classification organizes data in a hierarchical tree structure with root, internal, and terminal nodes, making decisions based on attribute selection measures. Naive Bayes Classification is a probabilistic model based on Bayes Theorem, suitable for high-dimensional data and efficient in text and medical classification. Rule-Based Classification uses IF-THEN rules for assigning class labels, offering high interpretability but facing challenges like rule conflicts and

scalability. Backpropagation Classification, a core of neural networks, optimizes model performance by iteratively reducing prediction errors using gradient descent. Lastly, Support Vector Machine (SVM) identifies an optimal hyperplane that separates classes with maximum margin, using kernel functions for handling non-linear data. These methods collectively support various applications from text analysis to medical diagnosis by enabling accurate and efficient classification.

2.1 Decision Tree Classification

Data mining techniques include generating classifiers as a technique for analyzing data [15]. There is an enormous amount of information that classification algorithms are capable of handling in data mining. The Decision tree classification algorithm is also useful for making predictions about categorical class names, labelling class names based on training sets, and classifying newly discovered data [16]. Structure of the decision tree classification contains root node, internal node and terminal node. This kind of structure is commonly used in tree data structures like binary trees, search trees, and decision trees. In these trees, internal nodes serve as decision or branching points, while leaf nodes represent outcome. The tree follows a hierarchical organization, where elements are arranged in parent-child relationships.

2.2 Naive Bayes Classification

Naive Bayes is a popular algorithm used in application such as text classification, spam detection, sentiment analysis, and disease prediction. It performs effectively on high-dimensional datasets by assuming feature independence. Known for its computational efficiency, it delivers good results even with small amounts of training data. The Naive Bayes classification algorithm is a probabilistic model built on the principles of Bayes Theorem.First, the prior probability of each target class label is calculated based on its occurrence in the dataset. Next, the probability of each attribute given a class label is determined. These probabilities are then used in Bayes Theorem to calculate the posterior probability for each class. The class with the highest probability is selected, and the input is classified accordingly. This method is especially effective for classification problems where the features are conditionally independent, making the Naive Bayes classifier well-suited for tasks such as text classification, spam filtering, and medical diagnosis.

2.3 Rule- Based Classification

Rule-based classification is a machine learning approach that classifies data based on a set of predefined or learned rules. These rules are typically in the form of IF-THEN statements, where the IF condition specifies a pattern in the input data, and the THEN part assigns a class label. Rule-based classification is highly interpretable and efficient, making it ideal for structured data with clear patterns. However, it may face challenges such as rule conflicts, scalability issues, and dependency on a well-labelled dataset. It is commonly implemented in decision trees, expert systems, and association rule mining for effective classification tasks.

The process starts with an empty rule set, indicating that no rules are initially defined. The algorithm analyzes patterns in the data to learn classification rules for each class. Each newly discovered rule is added to the existing rule set. This cycle repeats, with new rules continuously generated and incorporated, until no additional rules can be extracted. This approach is commonly used in rule-based learning systems, such as decision tree algorithms and association rule mining, where patterns are extracted to make classification decisions.

2.4 Backpropagation Classification

The Backpropagation (Backward Propagation of Errors) algorithm is a supervised learning algorithm used in training artificial neural networks. It is an optimization technique that adjusts the weights of a neural network by minimizing the error between predicted and actual outputs. Backpropagation, a fundamental technique in deep learning, relies on the gradient descent optimization algorithm to update model parameters.

Initially, the model identifies the error by calculating the variance between the predicted and true values. To improve accuracy, the model's parameters are updated to reduce this error. This process is repeated iteratively until the error reaches a minimum, ensuring optimal model performance. Once the error is minimized, the model is ready for classification, meaning it can accurately predict the correct class labels for new inputs. This iterative optimization approach is fundamental in training machine learning models, particularly in supervised learning algorithms like neural networks and gradient-based methods.

2.5 SVM Classification

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. Its ability to find an optimal hyperplane that best separates data points into different classes. Hyperplane Selection: SVM finds a decision boundary (hyperplane) that separates data points of different classes. Support Vectors: These are the data points closest to the hyperplane, which influence its position and orientation. Margin Maximization: The best hyperplane is the one that maximizes the margin between the two classes, ensuring better generalization. Kernel Trick (for Non-Linear Data): When data is not linearly separable, SVM uses kernel functions (e.g., polynomial, radial basis function) to map data into higher dimensions, making it easier to classify.

The first step involves finding a hyperplane that separates the data points of two different classes. To accomplish this, support vectors and margins are utilized to identify the optimal decision boundary that maximizes the separation between different classes. The hyperplane with the maximum margin is considered the optimal one, as it provides better generalization for unseen data. Finally, once the best hyperplane is identified, it is used to separate the dataset into distinct classes. SVM is widely used in applications like image recognition, text classification, and bioinformatics due to its effectiveness in handling high-dimensional data and ensuring robust classification.

3 COMPARISON OF CLASSIFICATION APPROACHES

The Table1. compares key supervised machine learning classifiers based on their strengths and limitations. Decision Tree Classifiers handle both categorical and numerical data but become complex and prone to overfitting with large datasets. Naive Bayes Classifiers are easy to implement but perform poorly on imbalanced data and lack feature selection capabilities. Rule-Based Classifiers work well on simple data but are hard to update for evolving datasets. The Backpropagation Classifier is a neural network model that's simple to program and automatically learns from data. It works well for complex problems but heavily relies on high-quality input data. If the data is poor, it may overfit and not perform well on new inputs. Support Vector Machines handle high-dimensional data effectively but require long training times, while Bayesian Pattern Classifiers are easy to program but highly dependent on data quality.

Table 1: Comparison of supervised machine learning classification techniques with their challenges

Algorithm	Features	Challenges
DTC (Decision Tree Classifier)	It classifies both categorical and numerical outcomes, but the attribute generated must be categorical.	Computational complexity increases with the addition of more training samples, leading to overfitting and challenges in model generalizability.
NBC (Naive Bayes Classifier)	It is easy to develop class label models, which are used for assigning class labels to problems.	Struggles with imbalanced data, leading to issues in data quality and feature selection.
RBC (Rule-Based Classification)	Efficient with basic data.	Challenges in modifying rules, affecting model generalizability and adaptability to complex datasets.
BPC (Backpropagation Classifier)	There is no need to learn special functions, and it is easy to program.	Highly dependent on input data, leading to potential overfitting and sensitivity to data quality.
SVM (Support Vector Machine)	Scales well with high-dimensional data and provides good results.	High computational complexity, making training time-consuming and affecting model scalability.

4 CONCLUSION

Supervised machine learning has emerged as a powerful approach for building predictive and diagnostic models, particularly in the healthcare sector. By leveraging labeled data, classification techniques such as Decision Trees, Support Vector Machines, Naive Bayes, and Backpropagation enable early and accurate disease detection. These methods offer substantial benefits in terms of efficiency and automation, yet they also face challenges related to data quality, model interpretability, scalability, and class imbalance. To fully realize the potential of machine learning in healthcare, future research should focus on developing more robust, explainable, and adaptable models. Addressing these challenges will be key to advancing machine learning-based healthcare solutions and ensuring their reliability and acceptance in the real world.

REFERENCES

[1] H. Jiang, B. P. Salmon, T. J. Gale, and P. A. Dargaville, "Prediction of bradycardia in preterm infants using artificial neural networks," *Mach. Learn. Appl.*, vol. 10, Dec. 2022.
[2] S. Bozyel et al., "Artificial intelligence-based clinical decision support systems in cardiovascular diseases," *Anatol. J. Cardiol.*, Feb. 2024.
[3] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics Med. Unlocked*, vol. 26, Jul. 2021.
[4] C. Boukhate, H. Y. Youssef, and A. B. Nassif, "Heart disease prediction using machine learning," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, Mar. 2022.
[5] K. M. M. Uddin, S. K. Dey, R. Ripa, N. Yeasmin, and N. Biswas, "Machine learning-based approach to the diagnosis of cardiovascular disease using a combined dataset," *Intell.-Based Med.*, vol. 7, May 2023.
[6] J. Rahman, A. Brankovic, and S. Khanna, "Machine learning model with output correction: Towards reliable bradycardia detection in neonates," *Comput. Biol. Med.*, 2024.
[7] D. Sandhya and R. Kamalraj, "Heart disease prediction using machine learning algorithms," *Int. Res. J. Eng. Technol. (IRJET)*, 2022.
[8] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques," *J. Electr. Syst. Inf. Technol.*, 2023.
[9] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *Int. J. Intell. Netw.*, vol. 3, pp. 58–73, Jun. 2022.
[10] M. H. K., "Heart attack analysis and prediction using SVM," *Int. J. Comput. Appl.*, vol. 183, no. 27, Sep. 2021.
[11] S. Asadi, S. E. Roshan, and M. W. Kattan, "Random forest swarm optimization-based heart disease diagnosis," *J. Biomed. Inform.*, vol. 115, Mar. 2021, Art. no. 103690.
[12] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, 2020, Art. no. 012072.
[13] A. Jamuna, "Survey on predictive analysis of diabetes disease using machine learning algorithms," *Int. J. Comput. Sci. Mobile Comput.*, vol. 9, no. 10, pp. 19–27, Oct. 2020.
[14] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthc. Anal.*, vol. 3, Nov. 2023, Art. no. 100130.
[15] R. Kumar and R. Verma, "Classification algorithms for data mining: A survey," *Int. J. Innovations Eng. Technol. (IJIET)*, vol. 1, no. 2, pp. 7–14, 2012.
[16] S. S. Nikam, "A comparative study of classification techniques in data mining algorithms," *Orient. J. Comput. Sci. Technol.*, vol. 8, no. 1, pp. 13–19, 2015.
[17] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management," *EURASIP J. Adv. Signal Process.*, 2021.
[18] Y.-Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry.*, vol. 27, pp. 130–135, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.
[19] B. Jijo and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol. Trends*, vol. 2, pp. 20–28, 2021.
[20] D. S. Char, M. D. Abràmoff, and C. Feudtner, "Identifying ethical considerations for machine learning healthcare applications," *Am. J. Bioeth.*, vol. 20, no. 11, pp. 7–17, 2020.
[21] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proc. ACM Int. Conf. Bioinf., Comput. Biol., Health Informatics*, 2018, pp. 559–560.

[22] A. H. Gonsalves, F. Thabtah, R. M. Mohammad, and G. Singh, "Prediction of coronary heart disease using machine learning," in *Proc. 3rd Int. Conf. Deep Learn. Technol. (ICDLT)*, 2019.
[23] J. Nourmohammadi-Khiarak et al., "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health Technol.*, vol. 10, pp. 667–678, 2020.
[24] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms," in *Proc. World Congr. Eng. Comput. Sci. (WCECS)*, vol. 2, San Francisco, USA, Oct. 2014.