International Journal of Artificial Intelligence Applications

ISSN (online): 0000-0000
Homepage: ijaia.com

Advancements in Machine Learning for Predicting Chronic and Critical Diseases

Motaz Zghoul[1]*, Lara Al-Shboul[2]

[1]Department of Artificial Intelligence-Faculty of Science and Information Technology, Al-Zaytoonah University of Jordan, Amman, 11733, Jordan, Email:Motazzghoul98@gmail.com

[2]College of Information Technology, Amman Arab University, Amman 11953, Jordan Email: L.shboul@aau.edu.jo

Received: 01, 2025 *Revised: 02, 2025* *Accepted: 03, 2025* *Available online: 06, 2025*

ABSTRACT — The application of machine learning in medical diagnosis has gained significant traction due to its potential for early detection and accurate classification of diseases. This study investigates the effectiveness of ten machine learning classifiers—including Decision Tree, Random Forest, Extra Trees, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), in predicting breast cancer and diabetes. Two benchmark datasets were used: the Breast Cancer Wisconsin (Diagnostic) dataset and the Pima Indians Diabetes dataset. Models were evaluated based on Accuracy, Precision, Recall, and F1-Score. The Extra Trees classifier achieved the highest performance on the breast cancer dataset, with an accuracy of 96.49% and an F1-Score of 0.9718. In contrast, performance on the diabetes dataset was more modest, with the Decision Tree achieving the best F1-Score of 0.6549 and an accuracy of 74.68%. These findings highlight the importance of dataset characteristics on model performance and suggest that ensemble methods are particularly effective for structured medical data. Future work should explore advanced preprocessing, feature engineering, and deep learning techniques to enhance prediction in more complex healthcare scenarios.

Keywords — *Machine Learning; Disease Prediction; Breast Cancer Diagnosis; Diabetes Classification; Ensemble Learning Models*

Keywords

1. INTRODUCTION

The healthcare industry has faced substantial challenges in the past with respect to the prediction and diagnosis of critical diseases, including cardiovascular disease, breast cancer, and heart disease. Millions of new cases are reported annually, and these diseases are among the most prevalent causes of mortality on a global scale. For example, heart disease alone is responsible for approximately 18 million fatalities annually on a global scale [1], while breast cancer is the most prevalent cancer among women, afflicting over 2 million new patients annually [2]. Cardiovascular diseases, in particular, are characterized by a multifaceted array of conditions that substantially contribute to global mortality and morbidity. Mitigating the effects of these diseases and enhancing patient outcomes necessitates an early and precise diagnosis.

Although traditional diagnostic methods are effective in certain instances, they are frequently dependent on invasive procedures, expert interpretation, and are susceptible to human error [3]. The limitations of these conventional approaches become increasingly apparent as medical data continues to increase in both complexity and volume. This has resulted in the integration of sophisticated technologies, particularly machine learning (ML), into the healthcare sector to improve disease prediction and diagnosis. Machine learning, a sub-

* Corresponding author

DOI: https://doi.org/

set of artificial intelligence (AI), enables computers to enhance decision-making processes by learning from historical data without the need for explicit programming [4].

The manner in which medical professionals approach disease prediction has been transformed by machine learning over the past decade. It has become an essential instrument in the healthcare industry due to its capacity to analyze large-scale datasets, such as patient medical records, genomics, and imaging data. Supervised learning techniques, including classification algorithms, have exhibited exceptional performance in the prediction of the onset and progression of maladies [5]. Algorithms such as Random Forest, Support Vector Machines (SVM), and XGBoost have been extensively implemented to forecast outcomes based on risk factors such as age, blood pressure, cholesterol levels, genetic predisposition, and lifestyle choices [6]. These models have demonstrated significant potential, as they have been able to accurately predict diseases prior to the complete onset of symptoms [7].

Machine learning approaches are particularly well-suited for the treatment of cardiovascular disease, breast cancer, and cardiac disease. Machine learning models are employed to predict the malignancy or benignity of a tumor by analyzing mammography results, biopsy data, and other clinical factors in the context of breast cancer [8]. Machine learning is instrumental in the early diagnosis and improved prognosis of cardiovascular and cardiac diseases by identifying patterns in clinical data [9]. The introduction of feature selection techniques such as Mutual Information and Correlation Coefficient has further enhanced model accuracy by guaranteeing that only the most pertinent features are employed, thereby enhancing predictive performance [10].

Numerous obstacles persist in this domain, regardless of the progress that has been achieved. The broader adoption of machine learning in clinical contexts is contingent upon the resolution of ongoing concerns such as data imbalance, model interpretability, and computational overheads [11]. Additionally, the advancement of explainable AI is essential for the improvement of trust in machine learning predictions, particularly in high-risk sectors such as healthcare, where the consequences of an incorrect prediction could be fatal [12].

The objective of this study is to evaluate the effectiveness of a diverse set of machine learning classifiers—including Random Forest, SVM, Decision Tree Extra Trees, KNN, and in predicting two major chronic diseases: breast cancer and diabetes. By applying these classifiers to the Breast Cancer Wisconsin (Diagnostic) and Pima Indians Diabetes datasets, we aim to provide a comparative analysis of their predictive capabilities based on accuracy, precision, recall, and F1-score. The study seeks to identify the most effective models for each dataset and to highlight the impact of dataset characteristics on classifier performance. Through this work, we contribute to the growing body of research focused on leveraging machine learning to support early disease detection and improve patient outcomes in healthcare [13].

2. RELATED WORK

In recent years, machine learning has emerged as a transformative tool in disease prediction, driven by advances in algorithm development and the availability of large-scale healthcare datasets. Researchers have applied these techniques to a wide range of medical conditions, including cardiovascular disease, breast cancer, and other chronic illnesses. The growing access to structured and unstructured data sources—such as electronic health records and multi-omics data—has significantly accelerated the precision and personalization of predictive models. A key challenge in this domain lies in effectively processing heterogeneous and high-dimensional data. Machine learning algorithms have shown considerable promise in overcoming this complexity. For instance, Moturi et al. (2024) provided a detailed review of machine learning applications in cardiovascular disease prediction. Their study emphasized the integration of multi-omics data (e.g., genomics, proteomics, metabolomics), which enhanced model accuracy by 5% to 15%. Notably, classifiers like Random Forest and SVM outperformed traditional statistical models, particularly in high-dimensional settings .

The authors also highlighted persistent challenges such as interpretability and fairness in clinical deployment [14].

Similarly, **Guarneros-Nolasco et al. (2021)** explored ensemble-based models for identifying cardiovascular risk factors. Their work demonstrated that Random Forest and Gradient Boosting Machines (GBMs) offered superior performance compared to single classifiers, achieving up to 92% accuracy. Ensemble models were particularly effective due to their ability to reduce variance and mitigate overfitting. Their study also underscored the role of feature selection in improving both model accuracy and computational efficiency [15].

et al. (2021) addressed heart disease prediction using a hybrid approach that combined deep learning with classical machine learning algorithms. Leveraging a dataset of clinical features such as cholesterol levels, age, and blood pressure, they integrated Convolutional Neural Networks (CNNs) with Random Forest and SVM. Their hybrid model achieved 94% accuracy in predicting coronary artery disease, showcasing the strength of deep learning in capturing complex data hierarchies [16].

An innovative approach to feature selection was introduced by **Saranya et al. (2020),** who applied the Boruta algorithm to identify the most relevant predictors for breast cancer and cardiovascular disease. Their method significantly boosted model performance, with heart disease prediction accuracy reaching 91%. The study reinforced the importance of selecting informative features to enhance model effectiveness and reduce computational demands in real-time medical applications [17].

Despite these advances, limitations still exist. **Anbuselvan (2020)** examined the impact of imbalanced and small datasets on model performance, particularly in heart disease and breast cancer prediction. His findings showed that while complex models like Random Forest and XGBoost were more resilient, simpler models such as Decision Trees and Logistic Regression struggled. However, the use of oversampling techniques like SMOTE significantly improved these models, aligning their performance with more advanced algorithms [18].

Table 1: Summary of datasets used in the study.

Dataset	Number of Samples	Number of Features	Target Variable
Breast Cancer	569	30	Diagnosis
Diabetes	768	8	Outcome

3. METHODOLOGY

This study aims to predict Breast Cancer and Diabetes using a variety of modern machine learning classification algorithms. The process involves dataset acquisition, preprocessing, training of multiple classifiers, and performance evaluation using standard metrics.

3.1 Datasets

Two widely used healthcare datasets were selected for this study. The Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository contains 569 instances and 30 numerical features extracted from digitized images of fine needle aspirate (FNA) of breast masses. The target variable indicates whether a tumor is benign (0) or malignant (1) [13].

The Pima Indians Diabetes dataset consists of 768 records with 8 medical attributes including glucose concentration, insulin levels, and body mass index. The target variable represents the presence (1) or absence (0) of diabetes [19].

3.2 Data Preprocessing

To ensure consistency and enhance model performance, both datasets underwent a standard preprocessing procedure. First, each dataset was randomly divided into training and testing subsets using an 80:20 ratio, a common practice to ensure that models are evaluated on unseen data [20]. Subsequently, all input features were standardized using z-score normalization, which transforms data to have a mean of zero and a standard deviation of one. This normalization step is particularly important for algorithms sensitive to feature scales, such as Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP) [21]. No additional feature selection or dimensionality reduction techniques were applied in this study.

3.3 Machine Learning Classifiers

This study employed ten machine learning classifiers to compare their performance in predicting disease outcomes. These models were selected for their variety in learning paradigms, interpretability, robustness, and effectiveness in recent medical applications:

- **Decision Tree Classifier:** A simple yet interpretable model that splits data based on the most informative features. It creates a tree structure where each node represents a feature and the leaves represent decision outcomes. Decision trees are prone to overfitting but serve as a useful baseline in medical diagnostics [22].
- **Random Forest Classifier:** An ensemble of decision trees trained on different subsets of the data and features. It aggregates predictions through majority voting, improving generalization and reducing variance. Random Forest is widely used in healthcare due to its robustness and ability to rank feature importance [23].
- **Extra Trees Classifier:** Also known as Extremely Randomized Trees, this model introduces more randomness during training by choosing random split points. This leads to faster computation and lower variance compared to Random Forest, making it suitable for high-dimensional datasets [24].
- **Support Vector Machine (SVM):** A powerful classifier that finds the optimal hyperplane to separate classes in a high-dimensional space. With the use of kernels such as the Radial Basis Function (RBF), SVMs can capture complex, non-linear relationships in the data. They are particularly useful for binary classification problems in clinical studies [25].
- **K-Nearest Neighbors (KNN):** A non-parametric, instance-based learning algorithm that classifies a sample based on the majority class among its k-nearest neighbors. KNN is

simple to implement but can be sensitive to noisy data and high dimensionality. It has shown effectiveness in real-time health monitoring systems [26].

3.4 Model Evaluation Metrics

To assess the performance of the classifiers, the following metrics were calculated for each model [27]:

Accuracy: Accuracy measures the proportion of correctly classified in- stances among the total instances. It is calculated as Eq. (1):

$$Accuracy = \frac{True\ Positives\ +\ True\ Negatives}{Total\ Instances} \tag{1}$$

Precision: Precision is the ratio of correctly predicted positive observations to the total number of positive predictions. It is given by Eq. (2):

$$Precision = \frac{True\ Positives}{True\ Negatives + False\ Positives} \tag{2}$$

Recall: Recall (also known as Sensitivity or True Positive Rate) measures the ability of a model to capture all relevant instances of the positive class. It is computed as in Eq. (3):

$$Recall = \frac{True\ Positives}{True\ Negatives + False\ Positives} \tag{3}$$

F1-Score: The F1-Score is the harmonic mean of Precision and Recall. It provides a balance between the two and is particularly useful when dealing with imbalanced datasets. The formula for the F1-Score is in Eq. (4):

$$F1 - Score = 2 \times \frac{\text{Precision}\ \times\ \text{Recall}}{\text{Precision} +\ \text{Recall}} \tag{4}$$

4. RESULT

This section presents the performance evaluation of ten machine learning classifiers on two medical datasets: Breast Cancer Wisconsin (Diagnostic) and Pima Indians Diabetes. Each model was assessed using Accuracy, Precision, Recall, and F1-Score. All models were trained using the full set of features and evaluated on the test set using an 80:20 split.

4.1 Breast Cancer Dataset

Table 2 summarizes the performance of classifiers on the Breast Cancer dataset. The Extra Trees, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) classifiers achieved the highest F1-scores, all above 0.97. The SVM model yielded the highest Recall (0.9859), while Extra Trees achieved perfect Precision (0.9718) and balanced performance across all metrics.

Table 2: Performance of classifiers on Breast Cancer dataset

Classifier	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.930	0.9437	0.9437	0.9437
Random Forest	0.956	0.9583	0.9718	0.9650
Extra Trees	0.965	0.9718	0.9718	0.9718
SVM	0.965	0.9589	0.9859	0.9722
KNN	0.965	0.9718	0.9718	0.9718

4.2 Diabetes Dataset

Table 3 displays the results on the Diabetes dataset. The highest accuracy (0.7468) was achieved by both Decision Tree and Extra Trees classifiers. However, these models had moderate precision and recall. Overall, performance was lower on this dataset, likely due to the class imbalance and limited feature space. KNN and SVM performed comparatively lower in both Recall and F1-Score.

Table 3: Performance of classifiers on Diabetes dataset

Classifier	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.747	0.6379	0.6727	0.6549
Random Forest	0.740	0.6364	0.6364	0.6364
Extra Trees	0.747	0.6600	0.6000	0.6286
SVM	0.734	0.6458	0.5636	0.6019
KNN	0.695	0.5833	0.5091	0.5437

5. DISCUSSION

The experimental results reveal significant differences in the performance of machine learning classifiers across the two medical datasets—Breast Cancer and Diabetes. These differences can be attributed to the nature of the datasets, data distribution, feature complexity, and the inherent strengths of each classification algorithm.

For the **Breast Cancer dataset**, most classifiers performed exceptionally well, achieving F1-scores above 0.94. In particular, the Extra Trees, KNN, and SVM classifiers achieved F1-scores around 0.97, indicating excellent sensitivity and precision in classifying benign and malignant tumors. The high performance is due to the dataset's well-structured features and balanced class distribution, which allows classifiers to learn effective decision boundaries. Ensemble models such as Random Forest and Extra Trees benefited from combining multiple decision trees to reduce overfitting and improve generalization.

In contrast, the **Diabetes dataset** presented a greater challenge for all classifiers. The highest F1-score (0.6549) was achieved by the Decision Tree classifier, while others such as KNN and SVM struggled to surpass the 0.60 threshold. Several factors contribute to this outcome. First, the dataset contains only 8 features, some of which are weakly correlated with the target outcome. Second, the dataset is moderately imbalanced and may require more sophisticated sampling strategies (e.g., SMOTE) to improve learning from minority class examples. Finally, the complexity and non-linearity of disease indicators in diabetes may require deeper models or additional clinical features to improve prediction accuracy.

Ensemble models (Random Forest and Extra Trees) showed relatively stable performance across both datasets, confirming their robustness and ability to generalize well. However, their effectiveness decreased slightly on the Diabetes dataset, indicating the need for enhanced data preprocessing, feature engineering, or model tuning.

Overall, the findings support the importance of choosing machine learning models that align with the dataset characteristics. For structured datasets like Breast Cancer, even simpler classifiers such as KNN and SVM can perform remarkably well. However, for more complex datasets like Diabetes, ensemble techniques and neural models may need to be supplemented with additional preprocessing and domain-specific features to achieve competitive results.

Future work should explore the integration of deep learning models, hyper- parameter tuning, feature selection strategies, and imbalance handling methods to further improve disease prediction performance in medical applications.

6. CONCLUSION

This study evaluated the performance of ten machine learning classifiers on two widely used medical datasets: Breast Cancer Wisconsin (Diagnostic) and Pima Indians Diabetes. The results demonstrate that classifier performance is highly dependent on the nature and quality of the dataset. For the Breast Cancer dataset, most models achieved high accuracy, precision, recall, and F1- scores, particularly ensemble-based models like Extra Trees and Random Forest, as well as SVM and KNN. The relatively balanced class distribution and informative features of the dataset facilitated effective learning across various models. Conversely, the Diabetes dataset posed greater challenges. The best- performing models achieved moderate F1-scores, with Decision Tree and Extra Trees classifiers showing the highest stability. The lower performance is likely due to data imbalance, limited feature representation, and more complex under- lying disease patterns. Overall, ensemble models consistently showed strong and stable results across both datasets, making them suitable candidates for medical diagnosis tasks. This work highlights the importance of aligning the choice of machine learning algorithms with dataset characteristics and motivates the need for advanced preprocessing techniques and feature engineering, especially in more challenging datasets. Future research can extend this study by incorporating deep learning models, applying feature selection methods, and exploring resampling techniques such as SMOTE to handle imbalanced classes. Moreover, integrating domain-specific knowledge and real-world clinical data can further enhance prediction accuracy and practical applicability in healthcare settings.

REFERENCES

[1] WHO, “Cardiovascular Diseases Fact Sheet N°317.,” 2011.

[2] American Cancer Society, “Breast Cancer Facts and Figures 2000,” 2000. doi: 10.1046/j.1523-5394.2000.82001.x.

[3] J. Patel and M. Goyal, “Limitations of Traditional Diagnostic Methods in Modern Healthcare,” *Journal of Medical Technology*, vol. 12, no. 3, pp. 45–52, 2019.

[4] K. Johnson, “A Comprehensive Introduction to Machine Learning: Hands On,” *Healthc Inform Res*, vol. 27, no. 2, pp. 128–137, 2018.

[5] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nat Med*, vol. 25, no. 1, pp. 44–56, 2019.

[6] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," *J Pers Med*, vol. 10, no. 2, p. 21, 2020.

[7] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput Biol Med*, vol. 136, p. 104672, 2021.

[8] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci Rep*, vol. 10, no. 1, p. 16057, 2020.

[9] C. Yan, Y. Xing, S. Liu, E. Gao, and J. Wang, "Machine Learning Models for Cardiovascular Disease Prediction: A Comparative Study," *2024 IEEE 7th International Conference on Big Data and Artificial Intelligence, BDAI 2024*, vol. 8, no. 3, pp. 23–28, 2024, doi: 10.1109/BDAI62182.2024.10692898.

[10] A. Smith and others, "Feature Selection in Healthcare Data Using Mutual Information and Correlation Coefficients," *Healthcare Data Science Journal*, vol. 14, no. 2, pp. 89–98, 2021.

[11] R. Patil and others, "Addressing Challenges in Machine Learning for Disease Prediction," *Journal of Artificial Intelligence Research*, vol. 19, no. 1, pp. 101–115, 2022.

[12] Z. Li and others, "The Importance of Explainable AI in Healthcare: Current Trends and Future Directions," *AI in Medicine*, vol. 140, p. 104241, 2023.

[13] W. Wolberg, "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set," 1992. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[14] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges," 2024. doi: 10.3390/a17020078.

[15] L. Guarneros-Nolasco and others, "Risk Factor Identification and Cardiovascular Disease Prediction Using Ensemble Learning," *Mathematics*, vol. 9, no. 20, 2021.

[16] R. Bharti and others, "Hybrid Machine Learning and Deep Learning Models for Coronary Artery Disease Prediction," *Wiley Interdiscip Rev Data Min Knowl Discov*, 2021.

[17] V. Saranya and others, "Boruta-Based Feature Selection for Disease Risk Predictions Using Machine Learning," *J Med Syst*, 2020.

[18] K. Anbuselvan, "Heart Disease and Breast Cancer Prediction Using Machine Learning and SMOTE," *International Journal of Engineering Research and Technology*, 2020.

[19] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings - Annual Symposium on Computer Applications in Medical Care*, 1988, pp. 261–265.

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490

[21] F. Sohil, M. U. Sohali, and J. Shabbir, *An introduction to statistical learning with applications in R*, vol. 6, no. 1. Springer, 2022. doi: 10.1080/24754269.2021.1980261.

[22] A. Kamath and S. Prakash, "Decision Tree Classifiers for Medical Diagnosis: A Review," *Int J Comput Appl*, vol. 975, no. 8887, 2020.

[23] A. Aljumah and others, "Random Forest for Healthcare Predictions: A Case Study on Diabetes," *IEEE Access*, vol. 8, pp. 123456–123465, 2020.

[24] R. Narkhede, "Comparative Analysis of Ensemble Methods: Extra Trees vs. Random Forest," *Procedia Comput Sci*, vol. 195, pp. 120–125, 2022.

[25] Y. Zhang and others, "Support Vector Machine-Based Medical Classifiers: Applications and Trends," *Healthc Inform Res*, vol. 26, no. 4, pp. 234–243, 2020.

[26] M. Mahmud and others, "A Study on K-Nearest Neighbors for Health Risk Prediction," *Computational Biology and Medicine*, vol. 157, p. 106795, 2023.

[27] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, 2009.