JCAI

Journal of Cognition and Artificial Intelligence (JCAI)
2025, Vol. 01, No. 02,PP 27-31
https://doi.org/10.33322/jcai.xxx

Early Disease Detection and Prediction using Machine Learning

Hafsa Nizami[1], and Syed Zain Ali Shah[2]

[1]*Sir Syed University of Engineering and Technology, Karachi, Pakistan*
[2]*Foreign Expert Faculty, Soochow University, Suzhou. Jiangsu China*

Correspondence Author: Hafsa Nizami (hnizami@ssuet.edu.pk)

Abstract

Early identification/detection of diseases is vital for improving patient outcomes but also reduces mortality in patients. The tremendous growth of healthcare data in the present day combined with advancements in computational power has dramatically increased the power and practicality of Machine Learning (ML) for modern day medical diagnostics. The incorporation of ML in the healthcare setting is altering the way we identify, treat, and cure illnesses. Identifying illness at an earlier stage, such as cancer, is paramount to decrease deaths and initiate treatment on time. This paper examines various ML algorithms, and their use in predicting and detecting diseases such as cancer, diabetes, and cardiovascular diseases, and compares the effectiveness of supervised versus unsupervised models through reviewing and discussing current case studies, on a variety of diseases. The study examines the challenges associated with implementation of these models as well as ethical concerns. It is concluded that Machine Learning could improve clinical decision making substantially when you are very thoughtful about the design of the model in regards to precision of the information, transparency to the decision making process, and fairness.

Index Terms: *Clinical Decision-Making, Electronic Health Records, Machine Learning, Models, and Prediction of Diseases.*

I. INTRODUCTION

In recent decades, evolving technology, particularly Artificial Intelligence (AI), has advanced rapidly across many industries, including the healthcare industry which has been most profoundly impacted. An impactful area of advancement is the use of AI, particularly Machine Learning (ML), in the healthcare diagnostics part of society for the early detection and monitoring of disease. Timely diagnosis is a significant contributor to an effective healthcare system, which generally leads to better treatment results, a reduction of financial burden, and improvement of the health of the person.

However, traditional methods of diagnosis are often resource-intensive, slow, and rely on subjective clinical judgment. Unlike traditional diagnostic processes, ML can provide a scalable, objective and data-driven type of diagnostic tool and can reveal hidden patterns and complex non-linear relationships in the medical datasets available today [1].

As aforementioned, machine learning (ML) is a subset of artificial intelligence (AI) that allows a system to learn from past information and subsequently make predictions or informed decisions without the need to be programmed explicitly for each task [2]. In the healthcare context, ML techniques can ingest large datasets or time-series data (e.g., Electronic Health Records (EHRs), imaging data or diagnoses, genomic sequences, and data streams from wearables/devices) to predict indicators or alerts of disease development. In fact, ML could identify risk indicators or red flags of disease even before symptoms surface. This ultimately paves the way for preventive medicine and alleviates burden on today's healthcare systems. Numerous ML algorithms have demonstrated strong performance in detecting conditions/diseases such as diabetes, cancer, cardiovascular disorders, Alzheimer's disease, and respiratory illnesses (like Covid-19). Approaches like Support Vector Machines, Decision Trees, K-Nearest Neighbors, and Deep Learning architectures have achieved high predictive accuracy as evident from research literature. Also, the developments in Natural Language Processing (NLP) allow the valuable information in unstructured text, such as physician comments, pathology reports, and other reports, to be extracted so that diagnostic accuracy can be enhanced.

The confluence of rising healthcare data availability, larger computational power, and decreasing storage and processing costs has enabled ML to become an invaluable aid to contemporary medicine/healthcare practice. Research worldwide, as well as regional institution contributions, points to its importance. For example, research at Sir Syed University in Karachi, Pakistan, has utilized neural networks and hybrid models to forecast cardiac risks and interpret diagnostic imaging, demonstrating the potential of local research in solving global healthcare problems [3].

However, implementing ML in clinical practice comes with some notable challenges. Concerns around—algorithmic bias, patient confidentiality, regulatory environments, and explainability, need to be addressed with careful effort towards ensuring ethical and safe implementation. Work/collaboration among clinicians,

data scientists, and policymakers will be pivotal to addressing these constraints.

Chronic diseases such as heart disease, cancer, diabetes, and arthritis, especially, pose critical challenges to healthcare systems since their diagnosis at an early stage usually dictates the success of the treatment and the long-term prognosis of patients.

As per a study by the World Health Organization (WHO), more than 70% of the deaths worldwide are caused by non-communicable diseases such as cancer and diabetes [4]. Early diagnosis can significantly improve survival rates, reduce healthcare costs, and saves time and efforts for all stake holders.

This paper explores the role of ML in early disease detection and prediction, and presents a comparative analysis of popular algorithms using various healthcare data dimensions, and highlights current applications and challenges. It also underscores the importance of localized research efforts in adopting AI technologies for better public health outcomes.

The rest of the paper is organized as; Section II provides an overview machine learning algorithms and their applications. Section III delves into the methodology. Section IV presents the performance comparison. Section V shows results and discussion. Section VI shows challenges and limitations. Section VII shows ethical and regulatory considerations. Section VIII is conclusion by summarizing key findings and finally Section IX suggesting future research directions.

II. MACHINE LEARNING ALGORITHMS AND APPLICATIONS IN HEALTHCARE

The key ML algorithms, i.e., Support Vector Machines (SVM), Random Forest, K-Nearest Neighbors (KNN), Neural Networks, and Logistic Regression, commonly used for early disease detection, their application areas, and typical diseases are represented in the Table I.

Table I: Machine Learning Algorithms and Their Medical Applications

S. No.	ML Algorithm	Primary Medical Applications	Use Case Example
1	Logistic Regression	Risk Prediction, Binary Classification	Predicting Heart Disease and Diabetes Likelihood—[5], [6].
2	Decision Tree	Rule-Based Decision Making in Diagnostics	Breast Cancer Diagnosis, Treatment Planning—[7].
3	Random Forest	Classification Using Structured Health Data	Predicting Hospital Readmission and Heart Disease—[5], [8].
4	Support Vector Machine (SVM)	High-Dimensional Classification and Medical Imaging	Tumor Classification, Parkinson's Disease Detection—[6], and [7].
5	K-Nearest Neighbors (KNN)	Patient Similarity Matching, Anomaly Detection	Diabetes Prediction, Rare Disease Detection—[9].
6	Gradient Boosting (GBM)	Complex Decision Modeling for High Accuracy	Cancer Recurrence and Chronic Kidney Disease Prediction—[5].
7	Neural Networks (MLP)	Pattern Recognition in Numeric Medical Data	Breast Cancer and Diabetes Classification—[10], and [5].

III. METHODOLOGY

This study follows a structured pipeline approach which is used to compare the effectiveness of various machine learning (ML) algorithms for early disease detection. It mainly consists of:

- Data Collection,
- Preprocessing,
- Feature Selection,
- Model Selection and Training, and
- Performance Evaluation.

A. Data Collection

We used three publicly available medical datasets from **UCI Machine Learning Repository** and **Kaggle**, which are widely accepted for benchmarking classification algorithms in healthcare research [6], and [8]:

- **Heart Disease Dataset (UCI)** – 303 records with features like age, sex, chest pain type, and cholesterol.
- **Pima Indian Diabetes Dataset (Kaggle)** – 768 instances with attributes such as glucose, insulin, BMI, and diabetes pedigree function.
- **Breast Cancer Wisconsin Dataset (UCI)** – 569 entries with 30 numeric features computed from digitized images of breast masses.

These datasets represent a diverse range of diagnostic problems for testing binary and multi-class classification algorithms.

B. Data Preprocessing

Preprocessing was applied to enhance data quality and improve model performance. Missing values were addressed using mean imputation or by removing sparse records. The continuous features were normalized by using Min-Max scaling to bring the values to [0, 1] range. The categorical variables such as gender or diagnosis labels were encoded by using one-hot encoding or label encoding as appropriate [7].

Irregularities were removed from selected data by using the Z-score technique, and the datasets were rebalanced where necessary using Category-based sampling to address class imbalance, which is common in medical data [5].

C. Feature Selection

To reduce the dimensionality and eliminate irrelevant features from data, we applied Recursive Feature Elimination (RFE) and Correlation Matrix Analysis. This

approach helped retain only the most significant attributes that contributed to model accuracy while avoiding multi-collinearity [9].

D. Model Selection and Training

We implemented a broad range of ML models by using **Python (v3.9)** and libraries such as **Scikit-learn**, **XGBoost**, and **TensorFlow** [7].
The models include:

- **Baseline Models**: Logistic Regression, Decision Tree, K-Nearest Neighbors, and Naive Bayes.
- **Ensemble Methods**: Random Forest, Gradient Boosting, AdaBoost, and XGBoost.
- **Neural Networks**: A shallow Multi-Layer Perceptron (MLP) and a Deep Neural Network (DNN) with 3 hidden layers.

Training was conducted using an **80/20 train-test split**, and a **10-fold cross-validation** was employed to ensure robustness. Hyperparameters were optimized using **Grid Search** and **Randomized Search**, adjusting factors such as learning rate, maximum tree depth, number of estimators, and neural activation functions [10].

E. Performance Evaluation

Each model was evaluated by using standard classification metrics:

- **Accuracy**: Overall proportion of correct predictions.
- **Precision**: The proportion/amount of true positive predictions among all positive predictions.
- **Recall** (Sensitivity): The proportion/amount of actual positive cases correctly predicted.
- **F1-Score**: Harmonic mean of precision and recall.
- **AUC-ROC**: Area under the receiver operating characteristic curve, useful for binary classification performance.

These metrics are essential in medical applications where minimizing false negatives is often more critical than maximizing accuracy alone [5], and [9].
Below in Figure 2, ML model of SVM algorithm is represented which shows Heart Disease Prediction with the help of Receiver Operating Characteristic (ROC) curve, showing model performance (AUC = 0.91 for SVM).

A ROC curve with an Area Under the Curve or AUC of 0.91 for a SVM model specifies excellent performance in heart disease prediction. This suggests that this model effectively distinguishes between individuals with and without heart disease, with a high degree of precision.

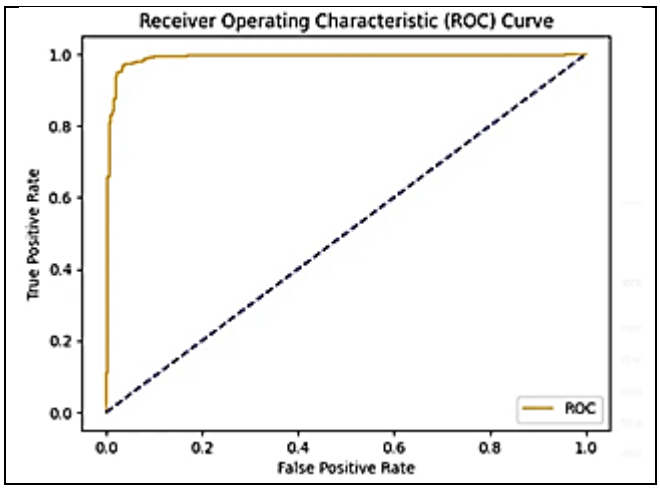

Figure 2: ML Model for Heart Disease Prediction (ROC curve showing model performance: AUC = 0.91 for SVM)

IV. PERFORMANCE COMPARISON

Early disease detection using machine learning requires rigorous comparison of models to identify those best suited for clinical deployment. Authors compares seven widely used models, i.e., Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), and Neural Networks (NN), across three key medical prediction datasets: Heart Disease, Diabetes, and Breast Cancer [6].
The Neural Networks or NN outperform all other models across all three datasets, achieving 92.3% accuracy on Heart Disease prediction, 83.5% on Diabetes, and 97.8% on Cancer Detection. Their superior performance largely stems from the capacity to model complex, nonlinear relationships within high-dimensional medical datasets [10].
Ensemble techniques such as Gradient Boosting and Random Forests have also shown strong predictive capabilities, particularly for structured data, due to their robustness and ability to minimize overfitting [5]. Support Vector Machines (SVM) often achieve competitive outcomes; however, their effectiveness is highly dependent on the careful selection of kernels and tuning of parameters [7].

Table III: Accuracy Comparison of ML Models on Disease Prediction Datasets

S. No.	ML Model	Heart Disease Accuracy (%)	Diabetes Accuracy (%)	Cancer Detection Accuracy (%)
1	Logistic Regression	85.4	78.2	94.1
2	Decision Tree	81.1	75.3	91.3
3	Random Forest	89.7	80.9	96.5
4	Support Vector Machine (SVM)	87.6	79.4	95.2
5	K-Nearest Neighbors (KNN)	84.5	76.1	93.5
6	Gradient Boosting	91.2	82.7	97.1
7	Neural Network	92.3	83.5	97.8

Note: Results averaged using 10-fold cross-validation. Neural Networks used a simple MLP with two hidden layers.

Although Logistic Regression is a comparatively simple linear model, it offers moderate predictive accuracy while remaining highly interpretable, making it particularly valuable in clinical settings where transparency is essential [5].

Decision Trees and KNN models perform relatively lower due to their tendency towards overfitting and sensitivity to noise.

The exceptional performance of Neural Networks, particularly when combined with explainable AI methods, makes them highly suitable for deployment in clinical decision support systems [9].

V. RESULTS AND DISCUSSION

The findings demonstrate that the selection of the modeling approach is dependent on purposes associated with data complexity and the requirements of clinical application. For example, neural networks perform well in situations where high-dimensional feature extraction is required [11] as in dermatological image analysis. Random forests are valid when the data is structured, as in laboratory test results or electronic health records [12].

KNN's lower performance in studies is attributed to the model's obvious dependence on irrelevant and noisy features [13]. KNN remains useful particularly in regard to clustering problems and exploratory analyses in cases examining varying clusters. Logistic regression models are interpretable compared to others and lack the same level of flexibility, making models with logistic regression applicable to examples assessing stroke and heart disease risk [14]. Developments in ensemble methods and hybrid architectures that combine different model types (for example CNNs and LSTMs) have emerged, indicating that these approaches may be useful potential solutions when examining multi-modal health data [15].

VI. CHALLENGES AND LIMITATIONS

Despite promising results, ML in healthcare faces significant challenges like:

- **Data Imbalance:** Rare conditions often lead to biased models.
- **Privacy Concerns:** Sensitive health data must be protected.
- **Interpretability:** Clinicians require understandable decision logic.
- **Generalizability:** Models trained on one dataset may not perform well on others.

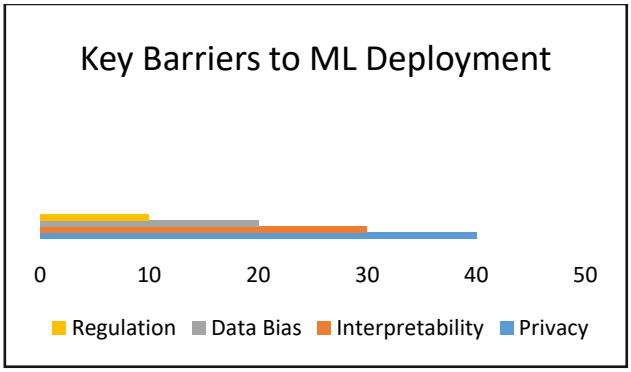

Figure 3: Key Barriers to ML Deployment in Clinical Settings

VII. ETHICAL AND REGULATORY CONSIDERATIONS

The following are ethical and regulatory considerations:

- **Bias Mitigation:** Addressing racial, gender, and age biases in datasets.
- **Transparency:** Using explainable AI (XAI) tools to improve clinician trust.
- **Regulation Compliance:** Adhering to GDPR, HIPAA, and FDA guidelines.

VIII. CONCLUSION

There is a great opportunity for machine learning to help with early detection of diseases. Our comparison showed that both neural networks and ensemble models such as Random Forest, provide good accuracy in predicting diseases. Nonetheless, if resources are limited, more simple models may be easier to interpret and may be preferred.

Machine learning presents a strong framework for early detection and prediction of diseases and represents a paradigm shift in prevent medicine and diagnostics.

While these models continue to improve in their accuracy and efficiency, the ethical issues, data quality issues, and transparency issues must be solved to allow for widespread clinical use.

IX. FUTURE RESEARCH DIRECTIONS

We recommend some future research directions as follows:

- Incorporating Real-Time Patient Data from Wearables (e.g., Fitbit, ECG monitors).
- Developing Explainable AI (XAI) Tools for Clinical Trust [16].
- Exploring Federated Learning to Maintain Patient Privacy [17].

Acknowledgment

The authors express their gratitude to the management of Sir Syed University of Engineering and Technology, Karachi, Pakistan, for their ongoing support and assistance during this study.

Authors Contributions

Each author made an equal contribution to this research study/work.

Conflict of Interest

The authors declare that there is no conflict of interest and affirm that this work is original, not plagiarized from any electronic or print source. All information drawn from external sources has been appropriately acknowledged and cited.

Data Availability Statement

The test data supporting this research are available in the paper.

Funding

No external funding was obtained for this research.

References

[1] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. Machine Learning for Healthcare Conference, 301–318.

[2] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

[3] Ahmed, S., Ahmed, E., Khan, A., & Rafiq, Z. (2022). Low cost and portable mechanical ventilator. Sir Syed University Research Journal of Engineering & Technology, 12(1), 57–63. https://sirsyeduniversity.edu.pk/ssurj/rj/index.php/ssurj/article/view/428.

[4] World Health Organization. (2021). Noncommunicable diseases. Retrieved from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[5] Mohapatra, D., & Rath, A. (2022). Comparative analysis of machine learning techniques for disease prediction. Journal of Biomedical Informatics, 128, 104027. https://doi.org/10.1016/j.jbi.2022.104027.

[6] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine. http://archive.ics.uci.edu/ml.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[8] Kaggle. (2023). Medical Datasets. Retrieved from https://www.kaggle.com/datasets.

[9] Hossain, M. S., & Muhammad, G. (2021). Explainable AI and mass surveillance system-based healthcare framework to combat COVID-19 like pandemics. IEEE Network, 35(1), 102–107.

[10] Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836.

[11] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. https://doi.org/10.1038/nature21056.

[12] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[13] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175–185.

[14] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE, 12(4), e0174944.

[15] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604.

[16] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[17] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19.