IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

A Review on the Role of Machine Learning Techniques in Early Detection and Diagnosis of Chronic Diseases

Sachin Yashwant Zurange[1] and Dr. Nilesh Vasant Ingale[2]
[1]Research Scholar, Department of Computer Science & Engineering
[2]Research Guide, Department of Computer Science & Engineering
Vikrant University, Gwalior (M.P.)

Abstract: *Chronic diseases such as cardiovascular diseases, diabetes, cancer, and chronic respiratory disorders are among the leading causes of morbidity and mortality worldwide. Early detection and accurate diagnosis are crucial for effective management and improved patient outcomes. Machine techniques have emerged as powerful tools for analyzing large and complex healthcare datasets, enabling predictive modeling and decision support systems in clinical practice. This review paper explores the role of machine learning techniques in the early detection and diagnosis of chronic diseases using clinical, imaging, and wearable device data. It discusses commonly used algorithms, data preprocessing methods, evaluation metrics, applications across different diseases, and associated challenges. A comparative table of machine learning techniques is also presented to highlight their strengths and limitations..*

Keywords*:* Chronic Diseases, Early Detection, Diagnosis, Predictive Analytics

I. INTRODUCTION

Chronic diseases develop over long periods and often remain undiagnosed until advanced stages, making early detection essential for effective intervention. Traditional diagnostic methods rely heavily on clinical expertise, laboratory tests, and imaging techniques, which may be time-consuming and sometimes limited in accuracy. With the rapid growth of electronic health records medical imaging, and wearable health technologies, vast amounts of clinical data are now available for analysis. Machine learning provides computational techniques that can learn patterns from these datasets and assist in early diagnosis and risk prediction (Rajkomar et al., 2019).

Chronic diseases represent a major global health burden, accounting for a significant proportion of mortality and long-term disability worldwide. Conditions such as cardiovascular diseases, diabetes mellitus, cancer, and chronic respiratory diseases develop gradually over time and often remain undetected in their early stages due to subtle or nonspecific symptoms. Early detection and timely diagnosis are therefore critical for effective intervention, improved patient outcomes, and reduced healthcare costs. Traditional diagnostic approaches rely heavily on clinical expertise, laboratory investigations, and imaging techniques, which, although effective, can be time-consuming, resource-intensive, and sometimes limited by human error or variability in interpretation. In this context, the integration of advanced computational techniques, particularly machine learning has emerged as a transformative approach in modern healthcare systems (Rajkomar et al., 2019).

MACHINE LEARNING TECHNIQUES USED IN CHRONIC DISEASE DIAGNOSIS

Machine learning techniques have become integral to the early detection and diagnosis of chronic diseases by enabling the analysis of complex, high-dimensional clinical data and uncovering hidden patterns that may not be easily identifiable through traditional statistical methods. Chronic diseases such as cardiovascular diseases, diabetes mellitus,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

cancer, and chronic respiratory disorders require timely identification to prevent complications and improve patient outcomes. Machine learning models leverage data from electronic health records laboratory test results, imaging modalities, and wearable devices to build predictive systems that assist clinicians in diagnosis and risk stratification.

Among the most commonly used approaches are supervised learning algorithms, unsupervised learning techniques, and deep learning models, each contributing uniquely to clinical decision support systems. Supervised learning algorithms, including logistic regression, support vector machines decision trees, random forests, and gradient boosting machines, are widely applied in disease classification tasks where labeled datasets are available. Logistic regression is often used for binary classification problems such as predicting the presence or absence of a disease due to its simplicity and interpretability, making it suitable for early screening applications. However, it assumes a linear relationship between input variables and the outcome, which may limit its performance in complex clinical scenarios.

In contrast, SVM is effective in handling high-dimensional datasets and works by identifying an optimal hyperplane that separates different disease classes. Kernel functions allow SVM to model nonlinear relationships, which is particularly useful in medical datasets where interactions among variables are not linearly separable (Cortes & Vapnik, 1995). Decision tree-based methods, such as random forests and gradient boosting, are also extensively used due to their ability to handle nonlinear relationships and mixed data types while providing relatively high interpretability.

Random forests, which consist of multiple decision trees trained on different subsets of the data, reduce overfitting and improve generalization by aggregating predictions through majority voting (Breiman, 2001). Gradient boosting methods, including XGBoost, build models sequentially by correcting errors from previous iterations, resulting in high predictive accuracy, especially in structured clinical datasets (Chen & Guestrin, 2016).

Unsupervised learning techniques also play a significant role in chronic disease diagnosis, particularly in identifying hidden patterns and patient subgroups within unlabeled datasets. Clustering algorithms such as K-means and hierarchical clustering are used to group patients with similar clinical characteristics, which can help in identifying disease phenotypes and stratifying risk levels.

Principal Component Analysis is another commonly used method for dimensionality reduction, which helps in simplifying large clinical datasets while preserving important variance. These techniques are particularly useful in exploratory data analysis and can support the development of more targeted diagnostic models (Hastie, Tibshirani, & Friedman, 2009). By uncovering latent structures in the data, unsupervised learning methods complement supervised approaches and enhance the understanding of disease heterogeneity.

Deep learning techniques have gained significant attention in recent years due to their ability to automatically extract hierarchical features from raw data without the need for manual feature engineering. Artificial neural networks form the foundation of deep learning and consist of interconnected layers of neurons that transform input data through nonlinear activation functions. Deep neural networks are particularly effective in modeling complex relationships in large-scale healthcare datasets. Convolutional neural networks are widely used for analyzing medical images such as X-rays, CT scans, and MRI images for the detection of diseases like cancer and cardiovascular abnormalities.

CNNs are capable of learning spatial hierarchies of features, making them highly effective in image classification tasks (LeCun, Bengio, & Hinton, 2015). Recurrent neural networks and long short-term memory networks are used for sequential or time-series clinical data, such as patient monitoring records and physiological signals, enabling prediction of disease progression and early warning of adverse events. These models are particularly valuable in intensive care settings where continuous monitoring data is available.

In addition to algorithm selection, the effectiveness of machine learning techniques in chronic disease diagnosis depends heavily on data preprocessing and feature engineering. Clinical datasets often contain missing values, inconsistencies, and noise, which must be addressed through data cleaning, imputation, normalization, and encoding of categorical variables. Feature selection techniques help identify the most relevant variables, reducing dimensionality and improving model performance.

Handling class imbalance is another critical step, as chronic disease datasets often have a disproportionate number of healthy and diseased cases. Techniques such as Synthetic Minority Oversampling Technique are commonly used to

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

balance datasets and prevent bias toward majority classes. Model evaluation is typically performed using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve with recall being particularly important in medical diagnosis to minimize false negatives and ensure that diseased patients are correctly identified (Saito & Rehmsmeier, 2015).

Machine learning models have been successfully applied in various chronic disease domains. In cardiovascular disease prediction, ensemble methods such as random forests and gradient boosting have demonstrated high accuracy in identifying patients at risk based on clinical parameters like blood pressure, cholesterol levels, and lifestyle factors. In diabetes diagnosis, machine learning algorithms are used to predict disease onset using features such as glucose levels, body mass index age, and family history. In oncology, deep learning models, particularly CNNs, have achieved remarkable success in detecting tumors from medical imaging data, enabling early cancer diagnosis and improved treatment planning.

Similarly, in chronic respiratory diseases such as chronic obstructive pulmonary disease machine learning models analyze spirometry data and environmental exposures to assess disease severity and progression. The integration of wearable device data, such as heart rate and physical activity levels, further enhances the ability of machine learning systems to monitor patients in real time and detect early signs of disease exacerbation.

Despite these advancements, several challenges remain in the application of machine learning for chronic disease diagnosis. One major issue is the lack of high-quality, standardized datasets, which can limit model generalizability across different populations and healthcare settings. Additionally, many complex models, particularly deep learning systems, function as “black boxes,” making it difficult for clinicians to interpret their predictions and trust their outputs. Efforts in explainable artificial intelligence are being developed to address this limitation by providing insights into model decision-making processes. Data privacy and security concerns also pose significant barriers to the widespread adoption of machine learning in healthcare, especially when dealing with sensitive patient information. Furthermore, integrating machine learning models into existing clinical workflows requires careful consideration of usability, interoperability, and regulatory compliance.

Machine learning techniques play a crucial role in the early detection and diagnosis of chronic diseases by enabling the analysis of large and complex clinical datasets. Supervised, unsupervised, and deep learning approaches each offer unique advantages in identifying disease patterns, predicting outcomes, and supporting clinical decision-making.

While challenges related to data quality, interpretability, and implementation persist, ongoing advancements in machine learning and artificial intelligence are expected to further enhance diagnostic accuracy and improve patient care. With continued research and collaboration between data scientists and healthcare professionals, machine learning has the potential to transform chronic disease management and contribute to more efficient, personalized, and proactive healthcare systems.

SUPERVISED LEARNING METHODS

Supervised learning algorithms are widely used when labeled datasets are available. These include:

Logistic Regression
Support Vector Machines
Decision Trees
Random Forest
Gradient Boosting
Artificial Neural Networks

These models are trained on historical patient data and used to predict disease outcomes such as the presence or absence of a chronic condition.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

UNSUPERVISED LEARNING METHODS

Unsupervised techniques are used for clustering and pattern discovery in unlabeled datasets:
K-Means Clustering
Hierarchical Clustering
Principal Component Analysis
These methods help identify patient subgroups and hidden structures within clinical data (Hastie, Tibshirani, & Friedman, 2009).

DEEP LEARNING TECHNIQUES

Deep learning models are particularly useful for complex and high-dimensional data such as medical images and time-series data:
Convolutional Neural Networks for imaging data
Recurrent Neural Networks and LSTM for sequential data
Deep Neural Networks for structured clinical data
These models automatically extract features without manual intervention and have shown high performance in disease detection tasks (LeCun, Bengio, & Hinton, 2015).

DATA SOURCES IN CHRONIC DISEASE PREDICTION

Machine learning models utilize diverse data sources, including:
Electronic Health Records
Laboratory test results
Medical imaging (X-rays, CT scans, MRI)
Wearable device data (heart rate, activity levels)
Genetic and genomic data
Integration of multimodal data improves prediction accuracy and supports comprehensive disease diagnosis.

DATA PREPROCESSING AND FEATURE ENGINEERING

Clinical datasets often contain missing values, noise, and inconsistencies. Preprocessing steps include:
Data cleaning and imputation
Normalization and scaling
Feature selection and extraction
Handling class imbalance (e.g., SMOTE)
Encoding categorical variables
Proper preprocessing enhances model performance and reliability (Johnson et al., 2018).

Table 1: Analysis of Machine Learning Techniques

Algorithm	Type	Advantages	Limitations	Applications in Chronic Disease
Logistic Regression	Supervised	Simple, interpretable	Limited nonlinear capability	Risk prediction (diabetes, CVD)
Support Vector Machine	Supervised	Effective in high dimensions	Computationally expensive	Cancer classification
Decision Tree	Supervised	Easy to interpret	Overfitting risk	Clinical decision support
Random Forest	Ensemble	High accuracy, robust	Less interpretable	Multi-disease prediction
Gradient Boosting (XGBoost)	Ensemble	High predictive performance	Requires tuning	Cardiovascular risk prediction
Artificial Neural	Deep	Captures nonlinear	Requires large data	Diabetes diagnosis

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

Network	Learning	relationships		
CNN	Deep Learning	Excellent for imaging data	Data-intensive	Cancer imaging detection
LSTM	Deep Learning	Handles time-series data	Complex training	Disease progression monitoring
K-Means Clustering	Unsupervised	Simple clustering	Requires predefined clusters	Patient segmentation
PCA	Unsupervised	Dimensionality reduction	Loss of interpretability	Feature reduction

APPLICATIONS IN CHRONIC DISEASE DETECTION

Machine learning techniques have been widely applied in the early detection and diagnosis of chronic diseases by leveraging large volumes of clinical, imaging, and physiological data to identify hidden patterns and predictive risk factors. These applications span across major chronic conditions such as cardiovascular diseases, diabetes mellitus, cancer, and chronic respiratory diseases, where early intervention is critical for reducing morbidity and mortality. By analyzing structured and unstructured healthcare data, machine learning models support clinicians in decision-making, risk stratification, and personalized treatment planning (Rajkomar et al., 2019).

In the domain of cardiovascular diseases machine learning algorithms are extensively used to predict the likelihood of heart attacks, coronary artery disease, and stroke. Clinical variables such as age, blood pressure, cholesterol levels, electrocardiogram readings, smoking status, and family history are commonly used as input features. Supervised learning models such as logistic regression, support vector machines, random forests, and gradient boosting machines have demonstrated high accuracy in classifying patients into high-risk and low-risk categories.

Ensemble methods like Random Forest and XGBoost are particularly effective in handling nonlinear relationships and interactions among clinical variables, making them suitable for cardiovascular risk prediction (Chen & Guestrin, 2016). These models enable early identification of at-risk individuals, allowing preventive interventions such as lifestyle modifications and medication management.

In diabetes mellitus, machine learning plays a significant role in both early diagnosis and disease progression monitoring. Predictive models are trained using patient data including fasting blood glucose levels, HbA1c, body mass index age, physical activity, and dietary habits. Logistic regression and artificial neural networks are commonly used for binary classification of diabetic and non-diabetic patients. More advanced models such as deep neural networks can capture complex nonlinear relationships between metabolic factors and disease onset.

Additionally, machine learning models are used to predict complications associated with diabetes, such as diabetic retinopathy and nephropathy, by analyzing retinal images and laboratory test results. These models enhance screening programs and facilitate early intervention strategies, thereby reducing long-term complications (Hastie, Tibshirani, & Friedman, 2009).

Cancer detection is another major application area where machine learning has shown remarkable success, particularly through the use of deep learning techniques. Convolutional neural networks are widely applied to medical imaging data such as mammograms, CT scans, MRI images, and histopathological slides for identifying tumors and classifying malignancies. These models automatically extract relevant features from raw images, eliminating the need for manual feature engineering.

For instance, CNN-based systems have been developed for early detection of breast cancer, lung cancer, and skin cancer with performance comparable to or even exceeding that of expert radiologists (LeCun, Bengio, & Hinton, 2015). In addition to imaging, machine learning models also analyze genomic and proteomic data to identify biomarkers associated with cancer progression and treatment response. These applications contribute to precision oncology, where treatment strategies are tailored to individual patients based on predictive analytics.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

Chronic respiratory diseases such as chronic obstructive pulmonary disease and asthma also benefit from machine learning-based diagnostic tools. These models utilize spirometry data, environmental exposure factors, patient symptoms, and medical history to predict disease presence and severity. Time-series models such as recurrent neural networks and long short-term memory networks are particularly useful for analyzing longitudinal patient data, enabling continuous monitoring of respiratory function. Machine learning algorithms can also detect exacerbations of respiratory diseases by analyzing wearable device data such as oxygen saturation levels and respiratory rate. Early detection of exacerbations allows timely medical intervention, reducing hospitalizations and improving patient outcomes (Topol, 2019).

In addition to disease-specific applications, machine learning techniques are widely used in electronic health record analysis for multi-disease prediction. EHRs contain a wealth of structured and unstructured data, including laboratory results, medication history, clinical notes, and diagnostic codes.

Machine learning models such as gradient boosting machines and deep learning architectures are trained on these datasets to predict the likelihood of multiple chronic diseases simultaneously. For example, predictive models can identify patients at risk of developing comorbid conditions such as diabetes and cardiovascular disease by analyzing shared risk factors and temporal patterns in patient records. This multi-label prediction capability enhances preventive healthcare strategies and supports integrated disease management (Johnson et al., 2018).

Wearable devices and remote monitoring systems have further expanded the applications of machine learning in chronic disease detection. Devices such as smartwatches and fitness trackers collect real-time physiological data including heart rate, physical activity, sleep patterns, and blood oxygen levels. Machine learning algorithms process this continuous stream of data to detect anomalies and predict potential health issues. For instance, abnormal heart rate patterns may indicate cardiovascular risk, while irregular glucose levels may signal diabetes-related complications. These technologies enable proactive healthcare by shifting the focus from reactive treatment to continuous monitoring and early detection.

Despite these advancements, the application of machine learning in chronic disease detection faces several challenges. Data quality and availability remain major concerns, as clinical datasets often contain missing values, inconsistencies, and biases. Additionally, the interpretability of complex models, particularly deep learning systems, poses difficulties for clinical adoption, as healthcare professionals require transparent and explainable predictions. Efforts in explainable artificial intelligence are being developed to address this issue by providing insights into model decision-making processes. Furthermore, ethical considerations such as patient privacy, data security, and algorithmic fairness must be carefully managed to ensure responsible deployment of machine learning systems in healthcare settings.

Machine learning techniques have a wide range of applications in the early detection and diagnosis of chronic diseases, spanning cardiovascular diseases, diabetes, cancer, and respiratory disorders. By leveraging diverse clinical datasets and advanced predictive models, these techniques enhance diagnostic accuracy, enable early intervention, and support personalized medicine. Continued advancements in algorithm development, data integration, and interpretability are expected to further strengthen the role of machine learning in chronic disease management and improve global healthcare outcomes.

CARDIOVASCULAR DISEASES

Machine learning models are widely used to predict heart disease risk using clinical parameters such as blood pressure, cholesterol levels, and ECG data. Ensemble methods like Random Forest and Gradient Boosting have shown high accuracy in identifying high-risk patients (Rajkomar et al., 2019).

DIABETES MELLITUS

ML models are used to predict the onset and progression of diabetes using glucose levels, BMI, age, and lifestyle factors. Logistic regression and ANN models are commonly used in early screening systems.

CANCER DETECTION

Deep learning models, particularly CNNs, are extensively used in analyzing medical images such as mammograms, CT scans, and histopathological slides for cancer detection (LeCun et al., 2015).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

CHRONIC RESPIRATORY DISEASES

ML techniques are applied to predict diseases like COPD and asthma using spirometry data, environmental factors, and patient history.

EVALUATION METRICS

Performance of ML models is evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- AUC-ROC

In chronic disease diagnosis, sensitivity is particularly important to minimize false negatives (Saito & Rehmsmeier, 2015).

II. CONCLUSION

Machine learning techniques have significantly advanced the early detection and diagnosis of chronic diseases by enabling the analysis of complex and large-scale clinical data. Supervised, unsupervised, and deep learning approaches each contribute uniquely to improving diagnostic accuracy and clinical decision-making. Despite challenges related to data quality, interpretability, and implementation, ongoing advancements in artificial intelligence are expected to enhance healthcare systems and support personalized medicine. With continued research and collaboration between clinicians and data scientists, machine learning will play an increasingly important role in combating chronic diseases globally.

REFERENCES

[1]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
[2]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference*, 785–794.
[3]. Hastie, T., Tibshirani, R., & Friedman, J. (2006). *The elements of statistical learning*. Springer.
[4]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.
[5]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.
[6]. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2017). MIMIC-III database. *Scientific Data*, 3, 160035.
[7]. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2018). MIMIC-III database. *Scientific Data*, 3, 160035.
[8]. LeCun, Y., Bengio, Y., & Hinton, G. (2016). Deep learning. *Nature*, 521(7553), 436–444.
[9]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
[10]. Rajkomar, A., Dean, J., & Kohane, I. (2017). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
[11]. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
[12]. Saito, T., & Rehmsmeier, M. (2015). Precision-recall analysis for imbalanced datasets. *PLoS ONE*, 10(3), e0118432.
[13]. Topol, E. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.