RESEARCH PAPER

Machine Learning Methods for Early Disease Detection Using Medical Data

Antima[1], Khel Prakash Jayant[2], Aishwarya Kashyap[3], Sudha Rani[4], Richa Mathur[5], Vibhanshu Tripathi[6]

[1]BCA Student, Department of MCA, Maharishi Markandeshwar (Deemed to be University) Mullana, Ambala (Haryana), India

Email: Antimamalik1856@gmail.com

[2]Professor, Department of Computer Science & Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, Uttar Pradesh, India

Email: drkpjayant@gmail.com

[3]Assistant Professor, Department of CSE(DS), Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India

Email: aishwaryakashyap42@gmail.com

[4]Assistant Professor, Department of School of Science (CS), Noida International University, Greater Noida, Uttar Pradesh, India

Email: bansalsudha903@gmail.com

[5]Associate Professor, Department of Computer Science and Engineering, Poornima University, Jaipur, Rajasthan

Email: richa.mathur@poornima.edu.in

[6]Assistant Professor, Department of Computer Science and Engineering, Ashoka Institute of Technology and Management, Varanasi, Uttar Pradesh, India

Email: vibhanshutripathi75@gmail.com

Received: 15th Jul, 2026 | Revised: 25th Jul, 2026 | Accepted: 5th Aug, 2026 | Available Online: 12th Aug, 2026

Abstract

A prompt diagnosis is crucial to better patient outcomes, lower health care costs, and prompt clinical action. Digital healthcare is booming, and with it, the amount of medical data available, such as Electronic Health Records (EHRs), laboratory reports, medical images, and data from wearable sensors, has grown significantly. In recent years, machine learning (ML) methods have proven useful for analysing these data and the patterns of disease that can be discovered by these methods that might not be easily discerned by traditional diagnostic approaches. The chapter summarizes some of the most prominent machine learning techniques applied for early disease detection such as: Supervised Learning algorithms (Logistic Regression, Decision Trees, Random Forest, Support Vector Machine) and Deep Learning models (Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks). It also introduces the medical data processing pipeline, highlights some of the major healthcare applications for diseases like cardiovascular disease, diabetes, cancer and chronic kidney disease, and explores some of the challenges currently faced in the medical landscape with respect to data quality, privacy and model interpretability. Moreover, novel research avenues like Explainable Artificial Intelligence (XAI), federated learning and multimodal analytics in healthcare are identified. In conclusion, machine learning has tremendous potential to improve early disease detection and enable intelligent healthcare systems that are driven by data, with the ultimate goal of advancing clinical decision-making and improving patient care.

Keywords: Machine Learning; Early Disease Detection; Medical Data; Artificial Intelligence; Healthcare Analytics; Deep Learning; Clinical Decision Support.

How to cite this article: Antima, Khel Prakash Jayant, Aishwarya Kashyap, Sudha Rani, Richa Mathur, Vibhanshu Tripathi. Machine Learning Methods for Early Disease Detection Using Medical Data. Int J Drug Deliv Technol. 2026;16(76s):1978-1987. DOI: 10.25258/ijddt.16.76s.189

Introduction

The key to better patient outcomes and cost savings and to avoiding the development of more severe diseases is to detect them early. The traditional techniques of disease diagnosis are mainly based on clinical skill, laboratory test and medical imaging methods, and sometimes these methods are time consuming and cannot detect diseases early enough. Digital healthcare systems have experienced rapid growth, leading to the creation of tremendous amounts of medical data, such as Electronic Health Records (EHRs), lab reports, medical images, and data from wearable sensors. These data offer crucial opportunities for the development of intelligent diagnostic systems that enable the timely and accurate detection of disease (Chen, Decary et al., 2024).

One of the significant applications of AI in healthcare is machine learning (ML), which is capable of learning patterns from past health information and giving predictions about disease outcomes with high accuracy. In the field of disease diagnosis, algorithms like Logistic Regression, Random Forest, Support Vector Machine, and deep-learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have shown to perform very well in diagnostics of diseases like heart disease, diabetes, cancer, and chronic kidney disease (Ahsan and Siddique, 2021a; Krishnan and Thandava Meganathan, 2026).

Although there have been great leaps, issues like lack of quality data, privacy concerns, the interpretability of models, and data imbalance remain to be tackled for the clinical implementation of machine learning systems. However, the continuous evolution of explainable AI, federated learning, and multimodal data analysis ensures that these technologies will be more and more precise and useful in the healthcare sector (Abbas, 2026).

This chapter summarizes the prominent machine learning techniques on early disease diagnosis from medical data. It covers the most popular algorithms, medical processing techniques, real-world healthcare uses and current problems and future research directions, offering a holistic view of how machine learning is being applied in the field of medicine today.

2. Background and Literature Review

The healthcare sector has been greatly affected by the swift digitization of health care, which has revolutionized healthcare data gathering, storage, and analysis. Today, healthcare professionals and institutions such as hospitals, diagnostic centers, research labs, wearable sensors, and mobile health apps generate vast amounts of electronic health data or EHRs, medical imaging systems, laboratory reports, genomic sequencing technologies, and wearable sensors. These databases provide valuable clinical information but, due to the complexity of these data, statistical methods used to accurately predict diseases have been overwhelmed. Machine learning has therefore become a formidable computational tool that can discover patterns and relationships in medical data that were previously unknown and can assist in the early detection of diseases with higher accuracy and efficiency (Chen, Decary et al., 2024).

In the last decade, AI in healthcare has seen tremendous advancement in the power of computers, cloud computing, big data analytics, and the rich collection of annotated medical data. The first computer-aided diagnosis systems were based on handcrafted rules, which were created by medical experts. These expert systems proved to be useful in specific clinical applications, but were not flexible and sometimes failed to perform well with new data. Machine learning models have overcome these shortcomings by using patient data in the past to learn and continuously improve prediction accuracy as new patient data is added (Mirbabaie, Stieglitz and Frick, 2021).

Medical datasets have a number of properties that make them different from typical datasets in other areas. They are generally composed of various components combining structured information like demographic data, lab readings, medication histories, and vital signs, as well as unstructured information like doctor notes, pathology reports, radiological images, and genomic sequences. Furthermore, healthcare data often exhibits various privacy-sensitive characteristics, such as measurement errors, missing values, class imbalance, and other data-related issues that make the analysis challenging. Thus, it is necessary to include detailed preprocessing methods that increase data quality prior to model development in effective machine learning systems (Intelligence-Based Medicine, 2026).

Electronic Health Records is one of the most promising sources of clinical information that can be used for disease prediction. EHRs store patient records across time such as medical treatments, laboratory findings, allergies, surgeries, physician notes, etc. These records allow machine learning models to make more detailed analyses of how diseases evolve over time, and not only based on a single clinical observation. In recent years, longitudinal data analysis has been especially useful in forecasting chronic diseases with a long latency period, such as diabetes, hypertension, and other diseases, so that health care workers can take preventive measures before the onset of irreversible disease (Chen, Decary et al., 2024).

Medical imaging is another domain that has showcased tremendous success in the field of machine learning. High dimensional image data, like X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound and mammography are produced by radiological modalities that demand expert interpretation. Over the last decades, deep learning models, especially Convolutional Neural Networks have made tremendous strides in automated image analysis, learning the visual hierarchy of features that correspond to pathological conditions. These models have been effective in detecting breast cancer, lung cancer, diabetic retinopathy, pneumonia, brain tumours and many other diseases with high diagnostic accuracy, decreasing diagnostic delays and helping the radiologist in clinical decision making (Esteva et al., 2019).

The use of wearable health devices and IoMT (Internet of Medical Things) technologies has also increased the possibility of early detection of disease by continuously measuring physiological parameters like heart rate, blood pressure, blood oxygen saturation, respiratory rate, body temperature, physical activity and sleep patterns. This differs from traditional hospital-based measurements, which are made in strictly controlled environments, because wearable sensors deliver health data in real-time in normal environment and can be used to detect small changes in the physiology of a patient before symptoms are evident that could be discovered by a machine learning algorithm. For chronic diseases, like cardiovascular, diabetes, and neurological disease, it is highly beneficial to be able to monitor them continuously (George, 2026).

Supervised learning algorithms have been applied in the diagnosis of disease in the literature, with many examples available. Logistic Regression is one of the most commonly used baseline models because of its simplicity, ease of computation, and interpretability. Decision Trees offer clear diagnostic rules that are easy to understand, and Random Forest algorithms use several decision trees to boost the power of the prediction and to make the prediction less prone to overfitting. In spite of the simplifying independence assumption, Naïve Bayes classifiers are also useful for probabilistic diagnosis in high dimensional biomedical datasets, while Support Vector Machines have shown good results in these types of biomedical datasets (Ahsan and Siddique, 2021b).

The complexity of medical datasets has grown, and so have the use of deep-learning methods. ANNs can be used for nonlinear modelling of clinical variables, RNNs and Long Short-Term Memory networks can be used for processing sequential healthcare data, like longitudinal EHR records, and physiological time-series. Recently, Transformer-based networks have garnered a lot of attention because of their capability to model long-range dependencies in massive clinical data. The advanced models have consistently given better performance than traditional ML algorithms in image classification, medical language processing, and multimodal healthcare analytics (Krishnan and Thandava Meganathan, 2026).

Multiple systematic reviews have noted the increasing impact of machine learning in a variety of medical fields. Tajidini and Kheiri (2023) reported significant enhancements in the accuracy of predictions using ensemble learning approach (Gradient Boosting, XGBoost, AdaBoost, and hybrid machine learning models). Likewise, Abbas (2026) highlighted the importance of combining feature selection, hyperparameter optimization, and EAI methods in boosting model reliability without compromising clinical interpretability. The advancements have led to greater acceptance of machine learning's use in real-world diagnostic environments.

Cardiovascular diseases are one of the most common causes of death in the world and machine learning has proven to be very successful in detecting cardiovascular diseases. Algorithms have been created based on demographic data, electrocardiogram signals, blood biomarkers, lifestyle factors, and medical history, and those algorithms have been proven to have high prediction accuracy that allows for identification of high-risk individuals before they begin to experience severe cardiovascular events. Some of the most common algorithms that have been reported to have high performance for heart disease prediction are Random Forest, SVM, and Gradient Boosting algorithms (Ahsan and Siddique, 2021b).

Machine learning has also greatly helped in the diagnosis of cancer. Deep learning techniques have come a long way in the fields of breast cancer classification based on mammographic images, lung cancer detection from computed tomography (CT) images, and skin cancer identification from dermoscopic photographs. The success of CNNs in medical image analysis has been attributed to their capacity to extract complex visual information from images without human interaction in the form of feature engineering. These features help to minimize inter-reliability in diagnosis and facilitate early intervention and better patient outcomes (Esteva et al., 2019).

Neurodegenerative diseases, such as Alzheimer's and Parkinson's, create a challenge for diagnosis, as signs of these diseases can be the same as those of normal ageing. Neuroimaging, cognitive testing, speech signals and genetic markers have been used to train machine learning models with promising results for differentiating between early stage disease and healthy individuals. These systems can enable prompt therapeutic interventions, which may help to reduce disease progression and enhance long-term health-related quality of life (Fauzi, 2026).

The COVID-19 pandemic has sped up worldwide research into machine learning-aided diagnosis. Many studies have led to the creation of predictive models based on chest radiography, CT scans, laboratory parameters and clinical signs, which allow for rapid identification of infected patients. In limited resource settings, deep learning-based methods showed high sensitivity and specificity, and significantly shorten the time to diagnosis compared to conventional laboratory methods. These research works have shown machine learning techniques adaptability to address the new development healthcare crisis (Mondal, Bharati and Podder, 2021).

Table 1. Summary of Recent Studies on Machine Learning for Early Disease Detection

Author(s)	**Disease/Application**	**Dataset Type**	**Machine Learning Method**	**Key Findings**	**Limitation**
Ahsan and Siddique (2021a)	General disease diagnosis	Clinical datasets	Logistic Regression, SVM, Random Forest	ML improves diagnostic accuracy across diseases	Limited real-world validation
Chen, Decary *et al.* (2024)	Early disease detection	Electronic Health Records	Deep Learning	Longitudinal EHRs improve early prediction	Data heterogeneity
Tajidini and Kheiri (2023)	Multiple diseases	Multisource medical data	Ensemble Learning	Hybrid models outperform single classifiers	Explainability challenges
Krishnan and Thandava Meganathan (2026)	Disease diagnosis	Medical imaging & clinical data	Deep Learning	CNNs and LSTMs achieve superior performance	High computational cost
Van de Putte and Speeckaert (2026)	Chronic Kidney Disease	Clinical records	Machine Learning Models	Early CKD detection significantly improved	Limited multicentre datasets

3. Machine Learning Methods for Early Disease Detection

The power of machine learning is that it can detect intricate patterns within vast medical datasets and make highly accurate predictions, making it one of the most influential technologies in the healthcare industry. Machine learning algorithms are different from traditional statistical models in that they continually enhance their predictive accuracy as they learn from past data. These methods allow doctors, nurses, and other medical staff to diagnose illnesses early on, tailor treatments to each patient's specific needs, minimize diagnostic mistakes, and better inform clinical decisions. The choice of machine learning algorithm is influenced by various factors such as the size of the data, the type of data, the complexity of its features, computational resources, interpretability needs, and the disease being studied (Ahsan and Siddique, 2021a).

3.1 Supervised Machine Learning

In the medical field, the datasets are frequently labelled with known disease outcomes, making supervised learning the most common machine learning paradigm used in healthcare. The algorithm is trained to recognize how the clinical features relate to the diagnostic labels, which it then uses to classify unseen patients. In the field of supervised learning, it is well suited for disease classification, disease prognosis prediction, and risk assessment.

Logistic Regression

Logistic Regression is one of the most straightforward and easily understood classification models that are applied in medical diagnosis. It predicts a patient's likelihood of being a member of a specific disease class with multiple clinical variables. Logistic regression, like Linear Regression, is widely used by healthcare researchers because the effects of various risk factors can be easily seen through the regression coefficients and odds ratios given by the regression model.

Decision Tree

Decision Trees recursively partition the dataset based on feature values that create the maximum amount of information gain. The resulting tree structure is similar to clinical decision making and it is easily interpretable by the health care professionals. The sequence of decision rules towards a given diagnosis can be easily followed by physicians, enhancing confidence in the automated diagnosis.

Random Forest

To overcome the deficiencies of single Decision Trees, Random Forest is an ensemble model that uses multiple Decision Trees. Randomly selected subsets of training samples and predictor variables are used for each tree, and the final prediction is determined by majority voting.

Support Vector Machine

In the medical domain, where data often consists of a large number of clinical parameters, Support Vector Machine (SVM) is well-suited for high dimensional data analysis. The algorithm determines an optimum decision boundary that best separates the different categories of disease with minimum error in the classification of diseases.

An SVM model can be implemented with different kernel functions (linear, polynomial, radial basis function), which can model complex nonlinear relationships commonly found in biomedical data. Many studies have been conducted on the superior performance of SVM in cancer diagnosis, gene expression analysis, neurological disorders detection and prediction of cardiac diseases. However, if large healthcare data sets are analyzed, there is a significant increase in computational complexity (Ahsan and Siddique, 2021a).

K-Nearest Neighbour

The K-Nearest Neighbour (KNN) algorithm classifies patients in accordance with the labels of their k nearest neighbours in the feature space. The disease characteristics of similar patients are expected to be similar, which is convenient for medical classification problems, and KNN is intuitive.

Naïve Bayes

Naïve Bayes classifiers are a type of probabilistic classifiers based on the Bayes' Theorem with the assumption of independence between predictor variables. While this is generally not true in clinical data, the algorithm is still very efficient and works remarkably well in many healthcare applications.

3.2 Unsupervised Machine Learning

Unsupervised learning is different from supervised learning in that it is used to analyze unlabelled medical data, without using the known disease categories. These algorithms are able to detect hidden structures, clusters, and associations in patient populations, which can be used to find disease subtypes that were not previously known, and to inform personalized medicine.

Clustering Algorithms

The clustering techniques like K-Means or Hierarchical Clustering are used to classify patients based on their clinical, laboratory or genetic characteristics. These analyses help to uncover phenotypes of the disease, patient stratification, and treatment response patterns.

Principal Component Analysis

Principal Component Analysis (PCA) is often used as a dimensionality reduction method prior to classification of diseases. The data in medical datasets are usually correlated and may have hundreds or thousands of variables, which makes the computation more complex and less efficient.

3.3 Ensemble Learning

Ensemble learning is the technique of making use of a number of machine learning models to boost the predictive performance compared to that of a single algorithm. Combining predictions from different classifiers tends to be more robust, and to have lower prediction errors, than a single classifier due to the different strengths and weaknesses of the classifiers.

Gradient Boosting

In Gradient Boosting the trees are built one after each other such that each tree fixes the errors made by the previous tree. The algorithm has been shown to have excellent performance in predicting chronic conditions from structured clinical data.

AdaBoost

Adaptive Boosting gives a higher weight to misclassified training examples in subsequent rounds of learning. This approach allows the model to progressively develop its accuracy in diagnosis by concentrating on the difficult cases.

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is one of the top-performing ensemble learning models that have been used in the field of healthcare analytics in recent years. Includes regularization methods, efficient parallel processing and automatic missing value processing.

Many comparative studies have shown that XGBoost is more successful in predicting cardiovascular disease, diabetes, kidney disease, liver disorders, and cancer compared to traditional machine learning algorithms because of its ability to detect complex nonlinear interactions between clinical variables (Tajidini and Kheiri, 2023).

3.4 Deep Learning Methods

Deep learning is a high-level form of machine learning that can automatically learn hierarchical representation of features from raw medical data without requiring a lot of manual feature engineering. The healthcare field has been revolutionized by the advent of high-quality GPUs and a wealth of annotated healthcare data, driving the adoption of deep learning in nearly every facet of healthcare diagnosis.

Artificial Neural Networks

Artificial Neural Networks (ANNs) are a collection of computational neurons connected in a network with three layers: the input layer, the hidden layer and the output layer. ANNs learn the relationships between clinical variables and disease outcomes in an iterative manner, which allows them to learn nonlinear relationships.

ANNs have been found to have high predictive power in the diagnosis of diabetes, heart disease, hypertension and chronic kidney disease with structured clinical data (Krishnan and Thandava Meganathan, 2026).

Convolutional Neural Networks

CNNs are networks with a particular focus on analyzing image data. CNNs automatically extract spatial features from radiological images using convolutional filters and pooling.

CNNs have been quite successful in diagnosis of breast cancer with mammograms, lung cancer with CT scan, diabetic retinopathy with retinal fundus images, skin cancer with dermoscopic images, and brain tumours with MRI images. They are often able to identify subtle visual abnormalities that are similar or equal to experts (Esteva et al., 2019).

Recurrent Neural Networks and Long short-term memory.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are tailored to handle sequential data within the healthcare sector, like longitudinal Electronic Health Records (EHRs), physiological monitoring signals, and wearable sensor measurements.

These architecture will be able to model temporal relationships between clinical observations, making it possible to predict disease progression, patient deterioration, hospital readmission and adverse clinical events early on. Unlike standard RNNs, LSTM networks do not suffer from the vanishing gradient problem, making them ideal for long-term health monitoring applications (Chen, Decary et al., 2024).

Transformer Models

New architectures for transformers are the most recent developments in the field of deep learning in healthcare. Unlike recurrent models, Transformers operate on entire sequences at once with self-attention, which enables them to effectively capture long-range dependencies.

Transformer models have been recently applied to Electronic Health Records, Clinical Language Processing, Genomic Analysis, and Multimodal Healthcare Systems. Future intelligent diagnostic systems (Applied Sciences Editorial Team, 2026) are likely to rely on these due to their scalability and outstanding prediction capabilities.

Table 2. Comparison of Major Machine Learning Methods for Early Disease Detection

Algorithm	Learning Type	Strengths	Common Medical Applications	Limitations
Logistic Regression	Supervised	Simple, interpretable, fast	Heart disease, diabetes	Assumes linear relationships
Decision Tree	Supervised	Easy to interpret	Clinical decision support	Overfitting
Random Forest	Ensemble	High accuracy, robust	Cancer, CKD, cardiovascular disease	Higher computational cost
Support Vector Machine	Supervised	Excellent for high-dimensional data	Cancer diagnosis, genomics	Slow on very large datasets
K-Nearest Neighbour	Supervised	Simple implementation	Diabetes, heart disease	Sensitive to dataset size
Naïve Bayes	Supervised	Fast and computationally efficient	Disease classification, clinical text	Independence assumption
XGBoost	Ensemble	High predictive accuracy	Multiple chronic diseases	Parameter tuning required
CNN	Deep Learning	Excellent image analysis	MRI, CT, X-ray diagnosis	Requires large labelled datasets
LSTM	Deep Learning	Handles sequential data	EHR analysis, patient monitoring	Computationally intensive
Transformer	Deep Learning	Captures long-range dependencies	Clinical NLP, multimodal healthcare	Very high computational requirements

4. Medical Data Processing Pipeline for Early Disease Detection

4.1 Medical Data Acquisition

The initial phase of the pipeline is to gather pertinent medical information from multiple health care resources. Structured and unstructured data are produced at massive scale by modern hospitals and healthcare organizations, such as Electronic Health Records (EHRs), Laboratory Information Systems (LIS), Picture

Archiving and Communication Systems (PACS), wearable sensors, mobile health applications and genomic sequencing platforms.

Structured data typically consist of patient information (age, gender, etc.), lab test results, medications, blood pressure, heart rate, blood glucose readings, and other numerical clinical data. Unstructured data include physician notes, discharge summaries, pathology reports, radiological images, electrocardiograms (ECGs), and genomic sequences. By combining these various data sources, the resulting comprehensive view of the patient's health can better capture the complexity of disease patterns, which in turn allows for more effective machine learning algorithms (Chen, Decary et al., 2024).

So clinical skill is needed to pick out real abnormalities from erroneous data entries.

Data Normalization

Medical variables often have different scales of numbers. Blood glucose for instance has a wide range of measurements, as does cholesterol, age, and blood pressure. Some algorithms like Support Vector Machines, K-Nearest Neighbour and Artificial Neural Networks work better when variables are normalized (min-max normalization or Z-score standardization).

Normalization helps variables with wide ranges not have too great an influence on the training of the model and helps to converge more quickly when optimizing.

Data Encoding

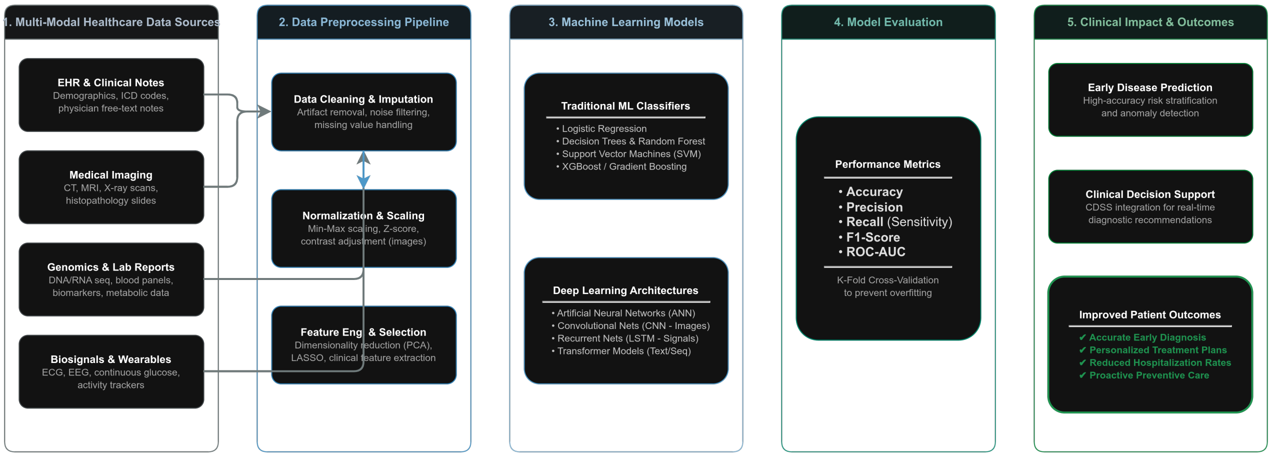


Figure 1. Overall workflow of machine learning-based early disease detection using medical data.

4.2 Data Preprocessing

Medical data is typically of a non-uniform nature that needs to be resolved before machine learning models can be effectively trained. Data preprocessing is hence one of the most important steps as it has a direct impact on the accuracy of the predictions.

Missing Value Handling

Laboratory data and/or patient data are often incomplete or missing from clinical datasets, as are diagnostic tests. The missing values can be due to a lack of certain examinations, lack of follow-up visits or data entry errors during clinical documentation.

Some of the most popular methods for dealing with missing data are mean imputation and median imputation for continuous data, mode imputation for categorical data, K-Nearest Neighbour imputation, multiple imputation, and model-based imputation. The choice of an imputation method ensures that the information is not lost from the data, and it also preserves the integrity of the data (Krishnan and Thandava Meganathan, 2026).

Noise Removal

There are many possible reasons for inaccurate observations in medical records, such as sensor failures, measurement errors in laboratories, transcribing errors, and sensor calibration errors. Pre-processing data with noise reduction methods like statistical filters, smoothing and outlier detection helps to enhance data quality prior to model training.

The detection of outliers is important since they can either be true disease conditions or a measurement error.

There are many health variables that do not have numerical values. Machine learning algorithms can then only be applied to these numbers and not to the disease stage, medication type, and blood group.

Depending on the nature of categorical variables, they can be encoded in various ways, such as Label Encoding, One-Hot Encoding, or Binary Encoding.

4.3 Feature Engineering

Feature engineering is a process of making informative predictor variables which make prediction of a disease better. The role of clinical knowledge is very important in finding meaningful combination of measurements in medicine that describe the disease.

For instance, body mass index (BMI) is derived from patient height and weight, and estimated Glomerular Filtration Rate (eGFR) is based on the patient's serum creatinine, age and gender to assess kidney function. In the same way, cardiovascular risk scores are a collection of several laboratory and demographic parameters combined into one single predictive parameter.

Feature engineering allows machine learning models to understand clinically relevant relationships that can't be intuitively understood from individual features (Abbas, 2026).

4.4 Feature Selection

In medical data, there are 100's or 1,000's of variables, which may be of little value in predicting the disease. Extraneous or redundant features can make the model more complex, make it more difficult to understand, and can cause overfitting.

Feature selection techniques select the variables that are the most predictive while removing irrelevant ones. Generally, these can be divided into three methods:

Filter Methods
Filter methods assess the variables without depending on the machine learning algorithm and employ statistical based methods like correlation coefficients, Chi-square tests, information gain and mutual information.
Wrapper Methods
Wrapper methods consider subsets of features and repeatedly train machine learning models to choose the subset which yields the best predictive performance. While being expensive, wrapper methods can yield very good feature subsets.
Embedded Methods
Embedded methods are those which select features while training the model. Estimating feature importance along with building predictive models is a natural process in Random Forest and XGBoost algorithms, and can be done simultaneously for feature selection and classification (Tajidini and Kheiri, 2023).
Clinically relevant features ensure a more efficient computation, more accurate predictions and more understandable models for healthcare professionals.

4.5 Model Training
After preprocessing and feature selection, the prepared data set is split into separate data sets for developing models.
One of the basic approaches is to divide the data into the following subsets:

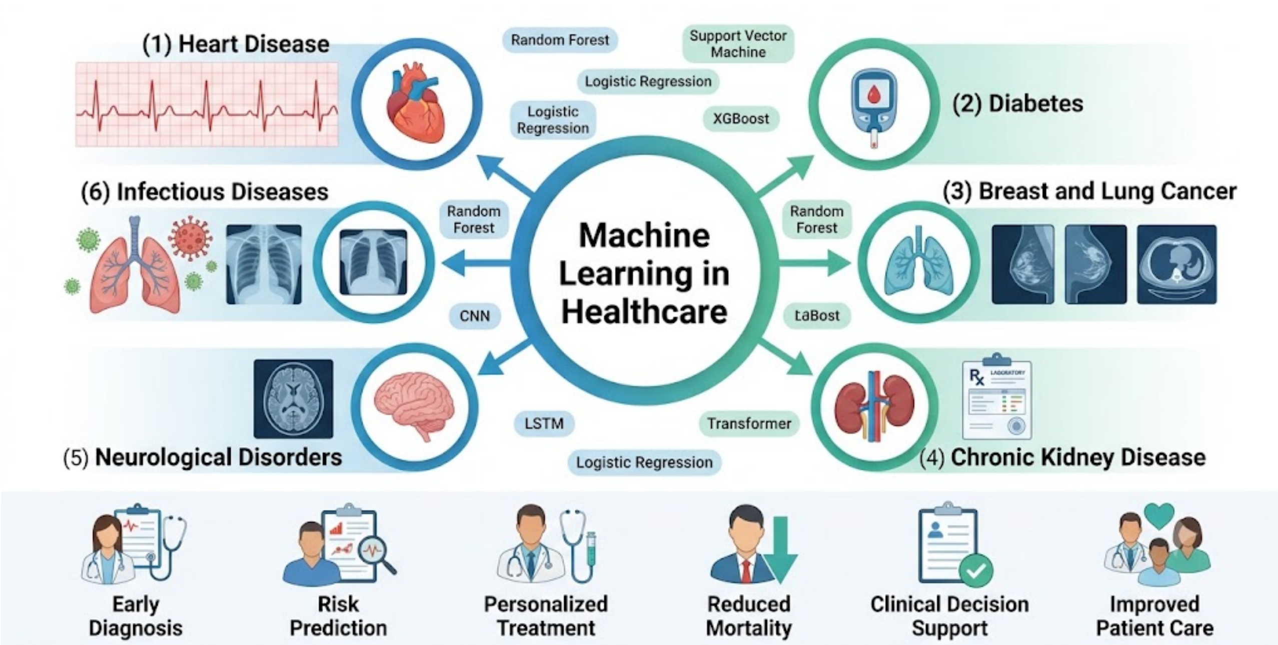
- Training set (approximately 70–80%)
- Validation set (10–15%)
- Testing set (10–20%)

The training set is used to learn relationships that exist between clinical variables and disease outcome. The validation set is used to fine-tune the model and its hyperparameters, and the testing set is used to assess the model's predictive power with new patient records.
Cross-validation (CV) methods, especially the k-fold CV method, are extensively used to ensure more reliable estimates of model performance and eliminate any bias due to small sample sizes (Ahsan and Siddique, 2021a).

4.6 Model Evaluation
Before machine learning models can be deemed appropriate for clinical use, they must be evaluated reliably. The accuracy and false diagnosis rates must be high for healthcare applications, since mispredictions can have direct consequences on patient safety.
There are a number of performance metrics used.
Accuracy
The accuracy is the percentage of correct patient classifications of all observations. While common, this may lead to misleading results if the classes of disease in the data set are highly imbalanced.
Precision
Precision is the percentage of people who have the disease who are correctly classified as having the disease. A high degree of precision reduces false positive diagnosis.
Recall (Sensitivity)
Recall: Percentage of patients with true positive predictions from the model. High sensitivity is especially important if early detection of disease is important because it may take longer to treat the affected patient if it is not caught.
Specificity
Specificity assesses the percentage of healthy people who are correctly classified healthy. High specificity means that few blood tests are ordered.
F1-Score
The F1 score, which is a combination of the precision and recall, is most useful when the classes of the various diseases are not evenly distributed and is especially relevant when there are few positives and many negatives.
ROC Curve (ROC)
The Receiver Operating Characteristic (ROC) curve represents the trade-off between the sensitivity and specificity for all possible classification thresholds.
Areas between curves.
The Area Under the ROC Curve (AUC) is a summary measure of the overall performance of the classifier. George (2026) suggests that values that are closer to 1 represent a good discrimination between healthy and diseased people.

Table 3. Medical Data Processing Pipeline for Machine Learning-Based Disease Detection

Pipeline Stage	**Main Activities**	**Purpose**	**Common Techniques**
Data Acquisition	Collect patient records, laboratory reports, medical images, EHRs	Gather comprehensive clinical information	EHR systems, PACS, wearable devices, laboratory databases
Data Preprocessing	Handle missing values, remove noise, normalize data, encode categorical variables	Improve data quality	Mean/median imputation, Z-score normalization, One-Hot Encoding
Feature Engineering	Create informative clinical variables	Enhance predictive capability	BMI calculation, eGFR estimation, risk score generation
Feature Selection	Select important predictors	Reduce dimensionality and overfitting	Chi-square, Information Gain, Random Forest Importance, XGBoost
Model Training	Train machine learning algorithms	Learn disease prediction patterns	Logistic Regression, SVM, Random Forest, CNN, LSTM
Model	Assess	Ensure model	Accuracy,



Evaluation	predictive performance	reliability	Precision, Recall, F1-score, ROC-AUC, Cross-validation
Clinical Deployment	Integrate model into healthcare systems	Support real-time diagnosis	Clinical Decision Support Systems (CDSS), Explainable AI, Continuous Monitoring

5. Applications of Machine Learning in Early Disease Detection

Machine learning is playing a crucial role in producing early forecasts of disease by utilizing vast amounts of healthcare information to pinpoint trends that might indicate a disease before it reaches a more critical stage. It helps healthcare practitioners to make faster and more accurate clinical decisions which will ultimately enhance patient outcomes and decrease healthcare costs (Rajkomar, Dean and Kohane, 2019).

It is also used extensively in the field of cardiovascular diseases, where algorithms process an electrocardiogram (ECG), blood pressure, cholesterol levels and patient history to help predict heart-related diseases in an early stage. Likewise, in diabetes detection, there are some clinical parameters like blood glucose, body mass index (BMI), and age are used to find the people who are at risk and will lead to diabetes so that they can be treated in time (Ahsan and Siddique, 2021b).

For example, Convolutional Neural Networks (CNNs) are a form of deep learning which can be used to identify tumors in medical imaging, such as breast and lung cancer. Machine learning can also be used to diagnose neurological diseases by analysing brain imaging, speech signals and cognitive evaluations, such as Alzheimer's and Parkinson's diseases (Esteva et al., 2019).

Figure 2. Major healthcare applications of machine learning for early disease detection.

6. Challenges and Future Research Directions

While machine learning has proven to be a valuable tool for early disease detection, it is not widely used in healthcare due to a number of challenges. In medical datasets, they may have incomplete data, imbalanced classes, and inconsistent information which can lead to a decrease in the accuracy of the models. Furthermore, issues of privacy, security, and interpretability of intricate deep learning models are considerable challenges to clinical deployment (Chen, Decary et al., 2024; Abbas, 2026).

Future work should consider the use of Explainable Artificial Intelligence (XAI), federated learning and multimodal data integration to develop more transparent and reliable machine learning models. Wearable technologies, Electronic Health Records, genomic information and large language models should help predict diseases and guide clinical choices more accurately. The model learning and continuous validation of models with the help of different healthcare datasets will further support the reliability and use of machine learning systems in regular medical practice (Gawali, Wang and Mukundan, 2026).

7. Conclusion

Machine learning is a potent technology that is increasingly able to detect disease early by analyzing medical information to better recognize disease patterns with high accuracy. Technologies like Logistic Regression, Random Forest, Support Vector Machine and deep learning models have shown great ability to diagnose the diseases like heart disease, diabetes, cancer and chronic kidney disease. Their capacity to process massive complex health information helps in early diagnosis, assists in clinical decision-making and enhances the outcomes for patients (Rajkomar, Dean and Kohane, 2019).

While data quality, privacy concerns, and model interpretability are important issues to tackle, the continued growth of the field of Explainable AI, federated learning, and multimodal healthcare analytics systems is improving the reliability and clinical relevance of machine learning systems. With the growing integration of digital technologies into healthcare, machine learning is likely to continue to be a key player in personalized medicine, patient prevention and proactive healthcare, and intelligent clinical decision support, thereby enhancing the efficiency and patient-centric approach to healthcare services (Gawali, Wang and Mukundan, 2026).

References

Abbas, F. (2026) ‘Machine-Learning-Based Disease Diagnosis and Prediction: Progress, Perspectives, and the Path Forward’, *Diagnostics*, 16(12), p. 1844.

Ahsan, M.M. and Siddique, Z. (2021a) *Machine learning based disease diagnosis: A comprehensive review*. arXiv preprint arXiv:2112.15538.

Ahsan, M.M. and Siddique, Z. (2021b) *Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review*. arXiv preprint arXiv:2112.06459.

Applied Sciences Editorial Team (2026) ‘Artificial Intelligence in Medical Diagnostics: Foundations, Clinical Applications, and Future Directions’, *Applied Sciences*, 16(2), p. 728.

Chaubey, P.K., Sharma, A., Turkani, R., Soni, P.K., Yadav, D. and Mishra, V. (2026) ‘Applications of Machine Learning in Early Disease Prediction: A Systematic Review’, *International Journal of Computer Artificial Intelligence*, 7(5), pp. 10–20.

Chen, M., Decary, M. and others (2024) ‘The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review’, *Journal of Medical Internet Research*, 26.

Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019) ‘A Guide to Deep Learning in Healthcare’, *Nature Medicine*, 25(1), pp. 24–29.

Fauzi, A. (2026) ‘Machine Learning for Early Disease Detection: A Systematic Review and Future Directions’, *Al-Kamal: Journal of Artificial Intelligence & Technology*, 1(1).

Gawali, S., Wang, X. and Mukundan, A. (2026) ‘Harnessing Machine Learning for Enhanced Biomedical Diagnosis and Early Disease Detection: Bridging Data Science and Healthcare’, *Frontiers in Oncology*.

George, E.A. (2026) ‘AI-Driven Healthcare Analytics for Early Disease Detection: A Scoping Review of Clinical Applications, Validation, and Translational Challenges’, *American Journal of Medicine and Health Studies*, 3, pp. 42–59.

Intelligence-Based Medicine (2026) ‘Medical Data Types in Machine Learning and Their Challenges: A Comprehensive Review’, *Intelligence-Based Medicine*, 15, p. 100427.

Krishnan, S. and Thandava Meganathan, N. (2026) ‘Disease Diagnosis and Prediction Using Deep Learning: A Review’, *PeerJ Computer Science*, 12, e3484.

Mirbabaie, M., Stieglitz, S. and Frick, N.R.J. (2021) ‘Artificial Intelligence in Disease Diagnostics: A Critical Review and Classification on the Current State of Research Guiding Future Direction’, *Health and Technology*, 11, pp. 693–731.

Mishra, V., Jayant, S., Dash, A., Chaudhary, C. and Babaria, S.S. (2026) ‘Artificial Intelligence and Machine Learning in Diagnostic Pathology: A Systematic Review of Applications, Challenges, and Clinical Implications’, *Cureus*, 18(3).

Mondal, M.R.H., Bharati, S. and Podder, P. (2021) *Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A Review*. arXiv preprint arXiv:2110.14910.

Rajkomar, A., Dean, J. and Kohane, I. (2019) ‘Machine Learning in Medicine’, *New England Journal of Medicine*, 380(14), pp. 1347–1358.

Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) ‘AI in Health and Medicine’, *Nature Medicine*, 28(1), pp. 31–38.

Tajidini, F. and Kheiri, M.J. (2023) *Recent Advancement in Disease Diagnostic Using Machine Learning: Systematic Survey of Decades, Comparisons, and Challenges*. arXiv preprint arXiv:2308.01319.

Topol, E.J. (2019) ‘High-Performance Medicine: The Convergence of Human and Artificial Intelligence’, *Nature Medicine*, 25(1), pp. 44–56.

Van de Putte, L. and Speeckaert, M.M. (2026) ‘Applications of Machine Learning for Early Diagnosis and Prognosis of Chronic Kidney Disease: Current Evidence’, *Diagnostics*, 16(15), p. 2354.