Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71119

# Development of Noble Method of Medical Diagnosis using Machine Learning Algorithms

Ambika Goyal
*Madhav Institute of Technology and Science*

***Abstract:** Machine Learning which is also known as  ML, a subset of Artificial Intelligence or called (AI), has been successfully used in the industry of healthcare to diagnose disorders. ML methods can diagnose both common and unusual diseases. However, the accuracy of machine learning in disease diagnosis remains an issue. The performance of different ML techniques varies depending on the healthcare dataset used. As a result, it is critical to apply numerous cutting-edge algorithms with excellent code efficiency in order to streamline the search for the best machine learning method for diagnosing a certain condition. ML is an autonomous system that learns by itself that has evolved from the use of explicitly coded data to deep analysis. One of the primary goal of the Machine learning system is to enable the machine to streamline the process and complete the task without the help of any  assistance. It incorporates and involves  human  envolvment within the analytic system.ML algorithms are divided into two categories: One is supervised and another is unsupervised. A supervised learning system is one in which the instances of the concerned data forecast its future data, however unsupervised learning is a procedure in which the data is taught, classified, or labeled. In our research, we show that various libraries can be used for the comparison of the performance of several machine learning algorithms for detecting a disease using a particular dataset, all with a few lines of code.  Medical diagnostics are performed with the goal of predicting illness identification with high accuracy and respecting each part of the human biological system. The identification of Cancer cells, heartbeat analysis, disease identification structure, evaluation of nervous system, ortho-care system, various disease identification analyses, and so on represent a few of the medical diagnosis applications. The applications of this kind of diagnostic system can provide the appropriate answer for speedy disease recovery.*

## I. INTRODUCTION

Machine Learning (machine learning), an aspect of artificial intelligence (AI), learns through data using various algorithms and is an iterative process that improves performance by making modifications along the way. ML has been successfully applied in nearly every field, including automation, education, travel, and health care. The process of machine learning techniques are most commonly used in healthcare to diagnose disorders. Machine learning approaches entered the health-care field in the 1970s, and the international AI journal Intelligent Computing in Medicine was launched in 1980. Over the next two decades, the illness diagnosis domain adopted conventional machine learning methods including supported vector machine learning, Naïve Bayes, and various artificial neural networks.  Additionally, investment in AI for applications in healthcare has increased significantly during the past ten years. Studies show that the application of artificial intelligence and machine learning (ML) in healthcare is resulting in the development of platforms, software, automated systems, and devices for health monitoring and enhancement.

Clinical data analysis can result in an early disease diagnosis, enabling the patient to start therapy on schedule. The conventional method of diagnosing illnesses is usually expensive and time-consuming. Additionally, the researchers demonstrate the time and cost effectiveness of machine learning-based illness detection systems.

Columns for different qualities and an additional variable for the classification variable are commonly found in a dataset tables used to construct a machine learning (ML) model for illness diagnosis. Whether or not the database instance has been positively detected for having the condition in question is indicated by the class variable. A positive diagnosis is usually indicated by a test result of 1, and a negative diagnosis is implied by a result of 0. Both supervised and unsupervised machine-learning algorithms were used to examine healthcare data. Generally speaking, supervised learning is the foundation of sickness diagnostic challenges. We will offer a thorough examination of the data collection and algorithmic learning techniques employed.

## II. PROBLEM STATEMENT

Although machine learning (ML) offers systematic and advanced methods for multivariate clinical data, its accuracy in illness detection remains a challenge. Furthermore, enhancing the performance about machine learning for diagnosing diseases is a hot issue in this field.

Because different ML approaches behave differently when applied to distinct healthcare datasets, we must find an effective way to apply many modern algorithms to the exact same data set in a reasonable amount of time and with few lines of code, allowing the search for the best ML method to be conducted efficiently to identify a specific disease.

Medical experts sometimes struggle to effectively diagnose diseases due to the complex structure of patient data, which might involve multiple variables and interactions.

Human errors in diagnosis can result in incorrect or delayed therapy.

Early identification is critical for effectively treating illnesses including diabetes, heart disease, and cancer.

## III. BASICS AND BACKGROUND

By analyzing data samples and drawing basic conclusions through statistical and mathematical methods, machine learning enables computers to learn without waiting for programming. This important accomplishment was first acknowledged in 1959 when Arthur Samuel published algorithms for pattern recognition in learning on experience and methods for prediction for video games. The fundamental objective of computational modeling. (ML) is to foresee or make judgment about a certain activity by learning from historical data.Many time-consuming jobs can now be completed swiftly and with little effort thanks to machine learning technology. Data-driven computational intelligence models that predict outcomes with near-perfect accuracy may now be trained thanks to the exponential growth in computing power and data capacity.

ML algorithms are often classified into three types: supervision, independence, and semi-supervised. However, algorithms used in ML can be divided into numerous subgroups determined by their learning methods. Linear regression, the logistic regression method, support vector machines (SVM), classification using random forest (RF), and naive bayes (NB) are some of the most popular machine learning approaches.

## IV. METHODOLOGY

The majority of chronic diseases are anticipated by our system. It accepts structured data as input for a machine learning model. End-users, such as patients or anyone else, use this system. In this method, the person using it will enter all of the symptoms that he or she is experiencing. These symptoms will subsequently be fed into the machine learning algorithm to predict the disease. Algorithms are then performed to determine which produces the highest accuracy. The system will then make disease predictions based on symptoms.

### A. *Data Collection*

Identify credible sources for health datasets, including hospitals and clinics with electronic health records (EHRs).

#### 1) *Source Identification*

Public health groups and databases owned by governments (e.g., CDC and WHO).

Institutions and academic search engines that publish health-related research.

#### 2) *Data types*

Include a range of data categories, such as demographics of the patient (age, gender, etc.).

Clinical signs and medical history.

Diagnostic test findings (such as blood testing and imaging).

Treatment results and follow-up data.

Structured data refers to tabular patient records that include both categorical and numerical characteristics.

Unstructured data includes X-rays, MRIs, and ECG signals for deep learning models.

#### 3) *Availability and Source. Reliability*

Open-source medical data sets can be accessed in the UCI Machine Learning Repository, including Heart Disease and Diabetes.

Kaggle (X-ray scans of pneumonia and breast cancer records).

MIMIC III (Electronic Health Records).

PhysioNet (ECG/EEG signals used in cardiovascular diagnostics).

#### 4) *Data Size and Quality*

A dataset must be sufficiently large to ensure generalization.

Imbalanced dataset (e.g., 90% well, 10% diseased) require balancing strategies.

*5) Ethical considerations*

Ensure compliance with ethical norms and legislation (such as HIPAA) governing patient privacy and data utilization.

### *B. Data Preprocessing*

Data preparation entails cleaning, manipulating, and arranging a dataset to enhance its quality before sending it into an algorithm for machine learning. This stage is critical for removing noise, addressing inconsistencies, and optimizing the dataset for training.

Data cleaning involves removing duplicate records to reduce bias in training models.Correct any errors in data entries (for example, normalizing units of measurement).

*1) Handling Missing Values*

Analyze the level of missing information and develop a strategy:

Imputation techniques (mean, median, mode, and more advanced techniques such as KNN imputation).

If imputation is not practicable, records with an excessive number of missing values should be removed.

*2) Outlier Detection and Removal*

Outliers are values that are exceptionally high or low, influencing model learning.

Techniques:

The Z-score Method identifies values that exceed a predefined threshold (e.g., ±3 standard deviation).

Interquartile Range (IQR): Eliminates numbers that are outside of the normal data range.

*3) Normalization/Standardization*

To ensure that features are on the same scale, use normalizing (scaling data to a range of [0, 1]) and standardization (growing data to a mean of 0 and a deviation from the mean of 1), which is critical for algorithms like SVM and KNN.

### *C. Feature Selection*

Feature selection is a critical stage in machine learning in general that involves determining the most appropriate variables which lead to reliable predictions while removing redundant or insignificant characteristics. Age, blood pressure, cholesterol levels, heart rate, and lifestyle factors are all taken into account when predicting heart disease. The study of correlation, mutual information, principal component evaluation (PCA), and recurrent feature elimination (RFE) are prominent techniques for dataset refinement. Effective feature selection boosts model performance, lowers overfitting, and increases interpretability, making predictions more dependable and therapeutically relevant.

1) Assess relevance by using statistical tests (e.g., the chi-square test for categorical data) to determine the link between characteristics and the variable being examined (diagnostic).
2) To minimize dimensionality, use techniques such as Principal Component Analysis ( PCA ) to retain critical information while minimizing the number of features.

Use algorithms like Random Forest to rank characteristics based on their predictive value for the target variable.

### *D. Data Splitting*

Training, Validation, and Testing Sets:

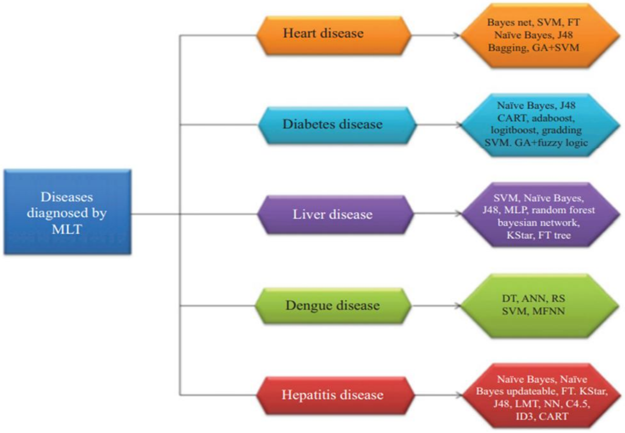
Separate the dataset into three pieces.

1) Training Set (70%): Utilized to train the model.
2) Validation Set (15%): Utilized to adjust hyperparameters and choose the best model.

The effectiveness of the final model on unidentified data is evaluated using the test set (15%).

### *E. Model Selection*

Choose the right machine learning algorithms depending on the task and the type of data. The most popular machine learning techniques for disease diagnosis are covered in detail in this section.


Diseases diagnosed by MLT
Heart disease
Bayes net, SVM, FT Naïve Bayes, J48 Bagging, GA+SVM
Diabetes disease
Naïve Bayes, J48 CART, adaboost, logitboost, gradding SVM. GA+fuzzy logic
Liver disease
SVM, Naïve Bayes, J48, MLP, random forest bayesian network, KStar, FT tree
Dengue disease
DT, ANN, RS SVM, MFNN
Hepatitis disease
Naïve Bayes, Naïve Bayes updateable, FT. KStar, J48, LMT, NN, C4.5, ID3, CART

Diseases and machine learning algorithm

*1) Decision Tree*

The divide and conquer strategy is used by the decision-tree algorithm. Classification trees are attributes that can have many values in DT models. While the branches symbolize the combination of attributes that lead to class labels, the leaves stand for distinct classes. Conversely, DT can be applied to continuous variables, such as regression trees. The two most popular and widely used DT algorithms are C4.5 and EC4.5.

*2) Support Vector Machine*

The support vectors machine, or SVM for short, is a widely used machine learning technique for regression and classification issues. Applications for SVMs include text categorization, remote homology discovery, diagnosis of diseases, tracking movements of the face, folded proteins, and recognized speech. Supervised machine learning methods cannot be used to examine unlabeled data. SVM clusters unlabeled data on a hyperplane to classify it. The SVM return, however, is not separable nonlinearly. When using SVM in data analysis, choosing the right kernels and characteristics are two crucial factors to address such problems.

*3) K-Nearest Neighbor (KNN)*

Evelyn Fix and Joseph Hodges developed this classification method in 1951 as a non-parametric approach. Classification and regression analysis can both benefit from KNN. Class membership is the end outcome of KNN classification. The object is categorized using voting techniques.The method used to determine the variation among each point or set of data points is called the Euclidean distance. The KNN values are averaged in regression analysis to determine the expected value.

*4) Naive Bayes*

One type of Bayesian classifier is Naive Bayes, or NB. Based on a particular record or piece of data, it determines the probability of acceptance in each class. The subcategory with the highest probability is the most likely.The NB classifier does not make predictions; instead, it evaluates likelihood.

*5) Logistic Regression.*

One machine learning technique for solving categorization issues is logical regression (LR). With expected results ranging from 0 to 1, the LR model uses a probabilistic framework. Online fraud detection, cancer detection, and spam email identification are a few instances of LR-based machine learning. LR uses the cost function, more commonly referred to as the sigmoid functions. Every real value is transformed between 0 and 1 by the help of sigmoid function.

*6) Artificial Neural Networks*

Multilayer perceptron-driven artificial neural systems are capable of expressing complicated non direct functions. AI relies heavily on artificial neural networks (ANNs). Seeing that the supplied label 'neural' suggests, it is cerebrum organized frameworks that are offered to imitate the method.

### *F. Model Training*

1) Training Process: The training process involves fitting selected models to a dataset to understand patterns and correlations.
2) Hyperparameter Tuning: Optimize model performance by tuning hyperparameters using strategies such as Grid Search or Random Search.
3) Cross-validation: Use k-fold cross-validation to ensure consistent model performance across subsets of training data.

## V. PERFORMANCE EVALUATION

The performance metrics used in the literature are covered in this section. When diagnosing illnesses, standard performance indicators include precision, recall, accuracy, and the f-1 score. If lung cancer is properly discovered, it can be categorized as either true positive (TP) or true negative (TN); if it is misdiagnosed, it can be classed as either false-positive (FP) or false negative (FN). These are the metrics that are most frequently utilized.

Accuracy is defined as the number of correct guesses divided by total number of forecasts. This is the primary metric used to assess the model. The formula is provided by

**Accuracy** =Tp+TN/Tp+TN+Fp+FN

Precision is the ratio of real positives over the sum of true positives plus false positives. It basically analyzes favorable forecasts.

**Precision**=Tp/Tp+Fp

Recall is the percentage of genuine positive cases that are appropriately identified.

**Recall**= Tp/Tn+Fp

Specificity is the percentage of genuine negative cases that are appropriately detected.

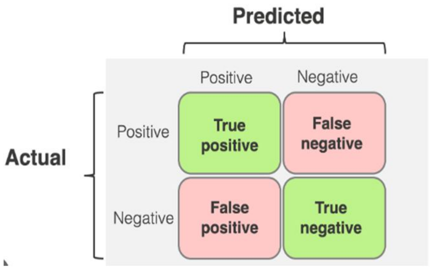
**Specificity**= TN/TN+FP

The harmonic average of memory and precision is represented by the F1 score. It offers one number that strikes a balance between recall and precision. It is particularly helpful when there is an unequal distribution of classes.

**F1 Score** = 2 x [(Precision x Recall)/ (Precision + Recall)]

The F1-score goes from 0 to 1,
where 1 denotes perfect recall and precision and 0 denotes neither.

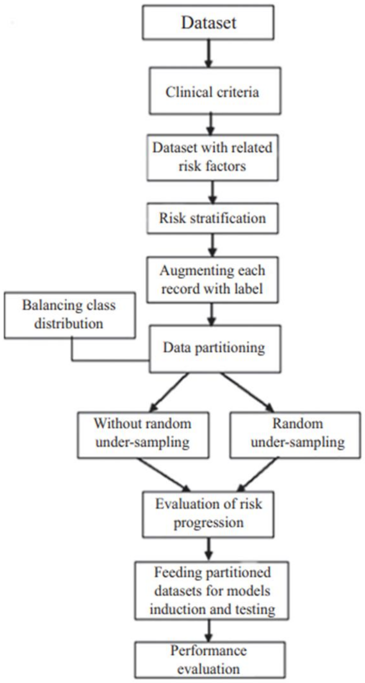
Confusion matrix:
This table displays the performance of a machine learning model. It is a tool that contrasts a dataset's expected and actual values.



|  |  | Predicted |  |
|---|---|---|---|
|  |  | Positive | Negative |
| Actual | Positive | True positive | False negative |
|  | Negative | False positive | True negative |

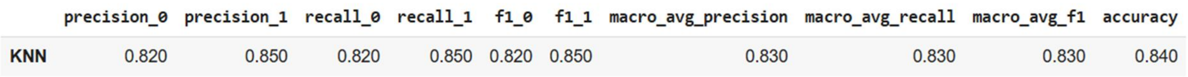
### *A. Heart Disease Prediction using Machine Learning*

Since cardiovascular disease is the leading cause of mortality globally, successful treatment and preventive measures depend on early detection. Due to its ability to evaluate intricate medical data and spot trends that more traditional methods might overlook, machine learning has become a powerful tool for heart disease prediction. Using health records from patients that contain demographic data, lifestyle factors, clinical signs, and medical history, machine learning algorithms can reliably forecast the risk of heart disease. Algorithms like logistic regression, decision tree algorithms, supports vector machines (SVM), as well as neural networks are commonly used to handle classification challenges. Following training, ML models are evaluated for reliability using metrics including precision, recall, precision, accuracy, and F1-score.


Dataset
Clinical criteria
Dataset with related risk factors
Risk stratification
Augmenting each record with label
Balancing class distribution
Data partitioning
Without random under-sampling
Random under-sampling
Evaluation of risk progression
Feeding partitioned datasets for models induction and testing
Performance evaluation

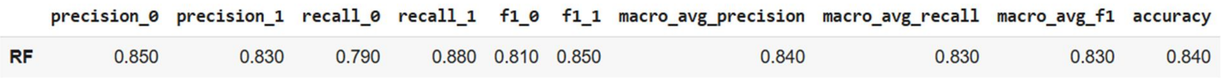
Steps to diagnose Heart disease using machine learning

Diagnosing Heart disease by KNN



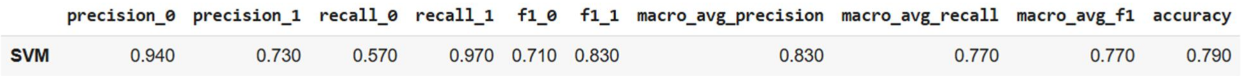
| | precision_0 | precision_1 | recall_0 | recall_1 | f1_0 | f1_1 | macro_avg_precision | macro_avg_recall | macro_avg_f1 | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| **KNN** | 0.820 | 0.850 | 0.820 | 0.850 | 0.820 | 0.850 | 0.830 | 0.830 | 0.830 | 0.840 |

Using Random forests



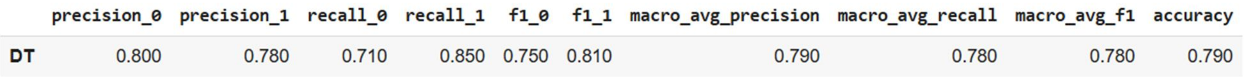
| | precision_0 | precision_1 | recall_0 | recall_1 | f1_0 | f1_1 | macro_avg_precision | macro_avg_recall | macro_avg_f1 | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| **RF** | 0.850 | 0.830 | 0.790 | 0.880 | 0.810 | 0.850 | 0.840 | 0.830 | 0.830 | 0.840 |

Using Support Vector Machines

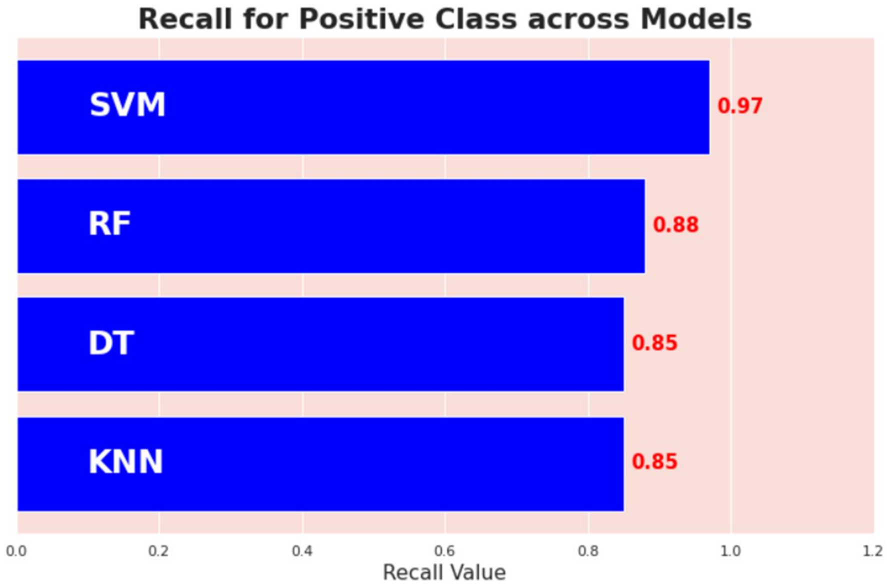


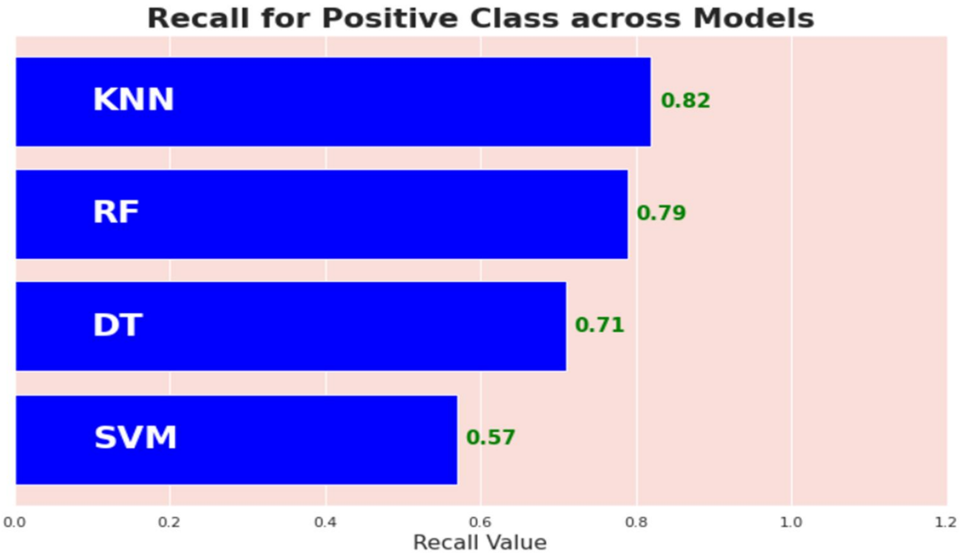
| | precision_0 | precision_1 | recall_0 | recall_1 | f1_0 | f1_1 | macro_avg_precision | macro_avg_recall | macro_avg_f1 | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| **SVM** | 0.940 | 0.730 | 0.570 | 0.970 | 0.710 | 0.830 | 0.830 | 0.770 | 0.770 | 0.790 |

Using DT



| | precision_0 | precision_1 | recall_0 | recall_1 | f1_0 | f1_1 | macro_avg_precision | macro_avg_recall | macro_avg_f1 | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| **DT** | 0.800 | 0.780 | 0.710 | 0.850 | 0.750 | 0.810 | 0.790 | 0.780 | 0.780 | 0.790 |





## VI. DISCUSSION

The use of machine learning in cardiac disease prediction represents a game-changing approach to early detection, utilizing massive volumes of patient data to improve clinical decision-making. ML models can examine risk factors and make accurate predictions using a variety of methods such as logistic regression, random forest models, decision trees, and neural networks. However, issues such as disparities in data, missing values, and potential biases in datasets must be carefully handled to ensure fair and reliable results. Furthermore, while ML models frequently show great accuracy, their black-box structure raises questions about interpretability and confidence in medical decision-making.

Explainable AI (XAI) approaches, such as SHAP values like LIME, can help close the gap by revealing how predictions are made. Furthermore, integrating machine learning with real-time tracking systems, portable health devices, and electronic medical records can improve continuous risk assessment and proactive response. Ethical considerations such as data protection, patient consent, and compliance with regulations are also important in the widespread implementation of machine learning-based diagnostics. Despite these obstacles, ongoing research and breakthroughs in artificial intelligence, ensemble modeling, and combined learning show promise for increasing predictive accuracy and therapeutic value. As machine learning advances, effective implementation in cardiac disease prediction has the potential to drastically reduce death rates and improve patient outcomes through early and individualized healthcare interventions.

## VII. FUTURE SCOPE AND CONSIDERATIONS

*1)* Deep learning advancements will increase heart disease prediction models' accuracy by finding complicated patterns in medical data.
*2)* Real-time monitoring via wearable devices will allow for ongoing monitoring of heart health and early intervention.
*3)* Explainable AI (XAI) improves model comprehensibility making Machine Learning (ML) predictions more reliable for healthcare practitioners.
*4)* Federated learning will allow for secure, global training of modeling across institutions while protecting patient privacy.
*5)* Improved data quality by removing values that are unavailable, noise, and bias ensures more accurate and fair forecasts.
*6)* Ethical and legal compliance with standards such as HIPAA and GDPR will be critical to ensuring openness and fairness.
*7)* Simple integration with electronic health records, or EHRs, will enable ML models to be used efficiently in clinical decision-making.
*8)* Personalized medicine will allow for customized treatment strategies based on an individual's genetics, lifestyle, as well as medical history.
*9)* Automated alarm systems powered by machine learning will send timely signals to physicians and patients about serious heart disease risks.
*10)* Continued research in the choice of features, model improvement, and false positive/negative detection will improve ML-based heart disease prediction.

## VIII. CONCLUSION

This Medical diagonsis system's primary objective is to forecast the illness based on its symptoms. This system generates a disease prognosis as its final output after receiving the symptoms of the client as input. This yields an average accurate forecast probability of 100%. The Grails framework was successfully used to build the Disease Predictor. This system is easy to use and offers a user-friendly environment. Anyone can access the system at any time and from any location. Lastly, in illness risk modeling, the variety of hospital data determines how good the risk prediction is. The use of machine learning (ML) algorithms for medical diagnosis has shown great promise for improving disease identification and prognosis. While classic models like DT, SVM, and NB continue functioning well, models using deep learning show great promise for medical imaging and difficult data analysis. ML-based diagnostic tools can improve early disease detection greatly when used in an organized manner that includes data preparation, feature selection, model training, and evaluation. The continued development of artificial intelligence (AI) healthcare solutions will alter the diagnostic procedures, making them simpler, more accurate, as well as available to a larger audience.

## REFERENCES

[1] Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives
Abstract: This paper explores the integration of AI algorithms in medical diagnostic systems, discussing their potential to enhance diagnostic accuracy and efficiency.
ieeexplore.ieee.org

[2] Healthcare Predictive Analytics Using Machine Learning and Deep Learning: A Comprehensive Survey
Abstract: This comprehensive survey reviews existing machine learning and deep learning approaches utilized in healthcare prediction, identifying inherent obstacles and future directions in the healthcare domain.
jesit.springeropen.com

[3] Machine-Learning-Based Disease Diagnosis: A Comprehensive Review
Abstract: This review explains how machine learning is being used to aid in the early identification of numerous diseases, addressing the challenges and complexities in developing early diagnosis tools and effective treatments.
mdpi.com

[4] A Study of Disease Diagnosis Using Machine Learning
Abstract: This paper discusses the application of machine learning algorithms in medical diagnosis, aiming to explore their impact on the accuracy and efficiency of disease diagnosis through case analysis and literature review.
mdpi.com

[5] Development of Machine Learning Classifiers for Blood-based Diagnosis and Prognosis of Suspected Acute Infections and Sepsis
Abstract: This study applied machine learning to the unmet medical need of rapid and accurate diagnosis and prognosis of acute infections and sepsis in emergency departments, achieving notable accuracy in disease diagnosis and severity prognosis.
arxiv.org

[6] Machine Learning Driven Biomarker Selection for Medical Diagnosis
Abstract: This research evaluates different methods for biomarker selection and machine learning classifiers for identifying correlations, aiming to improve the accuracy and efficiency of disease diagnosis.
arxiv.org

[7] Recent Advancement in Disease Diagnostic Using Machine Learning: Systematic Survey of Decades, Comparisons, and Challenges
Abstract: This review article examines machine-learning algorithms for detecting diseases, highlighting the collection of machine learning techniques and algorithms employed in studying conditions and the ensuing decision-making process.
arxiv.org

[8] Machine Learning Based Disease Diagnosis: A Comprehensive Review
Abstract: This review summarizes the most recent trends and approaches in machine learning-based disease diagnosis, considering factors such as algorithms, disease types, data types, applications, and evaluation metrics.
arxiv.org

[9] Machine Learning for Medical Diagnosis: History, State of the Art, and Perspective
Abstract: This paper provides an overview of the history, current state, and future perspectives of machine learning applications in medical diagnosis, discussing various algorithms and their effectiveness.
people.cs.rutgers.edu

[10] AI Breakthrough Raises Hopes for Better Cancer Diagnosis
Abstract: This article discusses a new AI foundation model that significantly advances the detection, prognosis, and treatment prediction of multiple cancers, achieving high accuracy across various cancer types.
ft.com