International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST

Survey On: Comparative Analysis of ML Models for Multi-Disease Prediction in Healthcare

Professor Sathish A[1], Purvika P[2], Shravani M R[2], Yashaswini N[2]

1Assistant Professor, Department of ISE, East Point College of Engineering and Technology, Bangalore, Karnataka, India

[2]Department of Information Science and Engineering, East Point College of Engineering and Technology, Bangalore, Karnataka, India

ARTICLEINFO

Article History:
Published : 30 May 2025

Publication Issue :
Volume 12, Issue 15
May-June-2025
Page Number :
186-192

ABSTRACT

This research investigates the use of machine learning (ML) models for the diagnosis of multiple diseases from medical datasets. Through the examination of large datasets with patient histories, symptoms, test results, and other health factors, ML algorithms are trained to identify diseases like diabetes, heart disease, and kidney disease. This paper compares several classifiers such as Random Forest, SVM, KNN, and deep learning models such as CNNs to evaluate their accuracy, precision, and recall in healthcare diagnosis.

I. INTRODUCTION

With increasing healthcare demands and large volumes of patient data, machine learning presents a powerful approach for disease prediction. Traditional diagnostic methods are often time-consuming and depend on specialist expertise, which may not be readily available in remote or resource-constrained settings. ML algorithms, trained on diverse healthcare datasets, provide a scalable, accurate, and cost-effective alternative. This paper focuses on identifying the best-performing algorithms for multi-disease prediction, considering the variability in symptom patterns, data imbalance, and feature relevance. Large-scale disaster management operations may struggle to scale up and deploy traditional disaster response procedures, which usually need labour-intensive and time-consuming manual coordination. Furthermore, these standard As a result of growing health care demands, there is an enormous amount of patient data available. This provides a new direction for automatic disease prediction using machine learning and artificial intelligence (AI) technologies. The current methods of diagnosis tend to be time- consuming and based on the presence of experienced individuals and experts that are usually lacking in rural communities or resource-scarce environments. The ML algorithms and approaches that have been trained on

heterogeneous health care datasets is, undoubtedly, a robust option that is helpful, accurate, and cost-effective.

II. LITERATURE REVIEW:

Over the past few years, incorporation of machine learning (ML) methods in healthcare has demonstrated encouraging results for diagnosing and anticipating several diseases at once. A number of studies have been directed at utilizing various ML algorithms for enhanced diagnostic accuracy, efficiency, and cost savings in multi-disease contexts.

Kavakiotis et al. [1] presented a detailed review of ML methods used in diabetes research and stressed the efficacy of support vector machines (SVM), decision trees (DT), and artificial neural networks (ANNs) in disease classification. Their results emphasized the importance of proper feature selection and pre-processing for enhancing prediction accuracy.

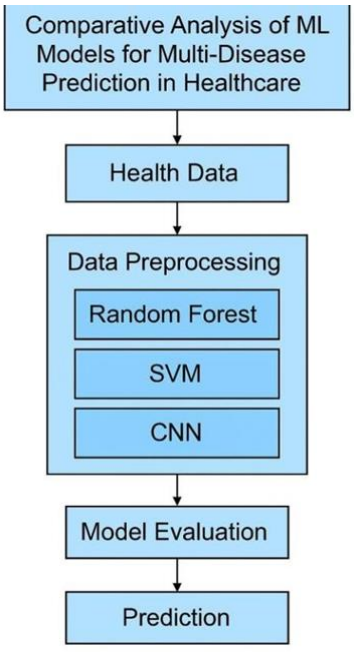
Choi et al. [2] presented a deep learning model with recurrent neural networks (RNNs) to predict multiple diseases from electronic health records (EHRs). Their study proved that RNNs perform better than conventional ML models in dealing with sequential data, specifically in temporal disease progression.

Another important study by Rajkomar et al compared different deep learning architectures on multiple hospital datasets. They found that deep models, when trained on large EHR data, have the ability to predict a wide range of diseases like heart failure, diabetes, and chronic kidney disease with high accuracy but also emphasized the importance of model interpretability in the clinical environment.

Shickel et [3] investigated the difficulties in implementing deep-learning in healthcare, observing that despite the success of models such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, data scarcity, imbalance, and limited availability of labelled data present challenges to their practical use

III. SYSTEM ARCHITECTURE:

Diagram:



Health Data Input Sources:

Electronic Health Records (EHR), laboratory tests,

Wearables sensors, imaging reports (X-rays, MRIs)

1. **Data Types:**

Structured (e.g., demographics, laboratory test results) and unstructured (e.g., images, clinical notes)

2. **Preprocessing Cleaning:**

Dealing with missing data, noise Normalization: Bounding numeric features into a uniform range Feature Engineering: Adding new features (e.g., BMI from weight and height) Dimensionality Reduction: Methods like PCA or RFE for optimizing model efficiency

3. **ML Models**

Random Forest (RF): Suitable for tabular clinical data; interpretable Support Vector Machine (SVM): Suitable for binary classification; feature scaling necessary Convolutional Neural Network (CNN): Most suitable for image data such as X- rays or MRIs Each model is trained on the preprocessed dataset and cross-validation is used to tune them.

4. **Model Evaluation**

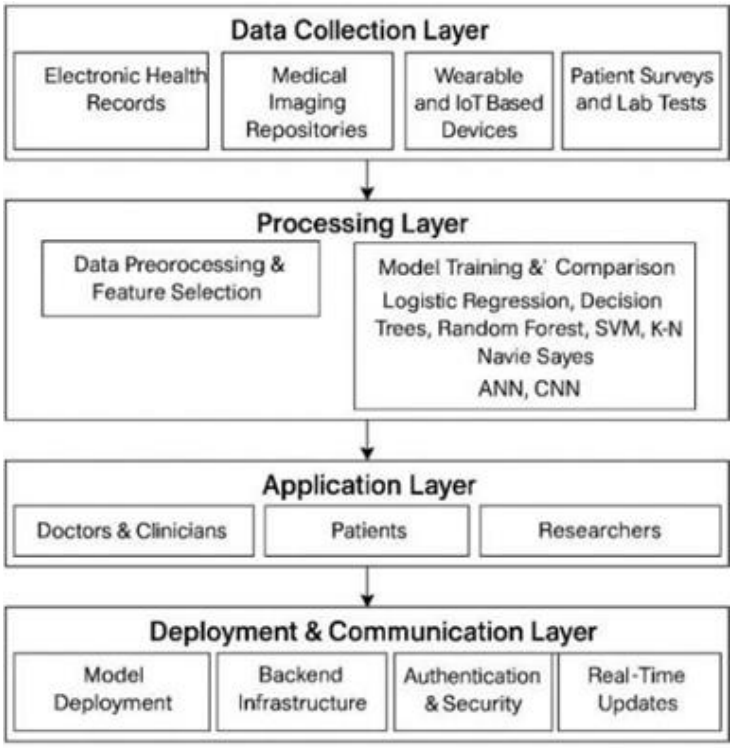
Accuracy Precision & Recall F1 Score ROC-AUC (for binary classifiers) validation Methods: k-fold cross-validation, holdout validation

5. **Prediction Output**

Disease risk score, binary classification (Yes/No), or ranked condition list

Overall Architecture:

There are four primary layers



1. **Data Collection Layer**

Collects patient information from EHRs, wearable sensors, lab tests, and public data for the training of ML models.

2. **Processing Layer**

Pre-processes and cleans data, extracts features, and trains different ML models (such as SVM, Random Forest, ANN) to predict diseases.

3. **Application Layer**

Delivers user interfaces to doctors, patients, and researchers to view predictions, insights, and visualizations.

4. **Deployment & Communication Layer** Houses the models and applications securely, supporting real-time prediction and guaranteeing system access, scalability, and data privacy

IV. METHODOLOGY:

Collecting and Preparing Data:

The framework collects healthcare data from several credible sources including: UCI Machine Learning Repository Kaggle medical datasets Public health databases (e.g., CDC, WHO)

Data types generally consist of:

Patient demographics (age, gender, etc.) Medical history and symptoms Laboratory test results and vital signs\ Disease diagnosis labels (for diabetes, heart disease, Parkinson's, etc.)

Data preparation consists of:

Managing missing and inconsistent values scaling continuous features Encoding categorical data Managing class imbalance using algorithms such as SMOTE

Feature Selection and Extraction

Relevant features are chosen by:

Correlation analysis

Correlation analysis Recursive Feature Elimination (RFE) Principal Component Analysis (PCA) This reduces dimensionality and improves model performance by removing irrelevant or redundant information.

Model Training and Development

Several machine learning models are trained to predict disease probability: Logistic Regression – For binary/multi class disease classification K- Nearest Neighbours (KNN) – For instance- based learning Support Vector Machine (SVM) –For best separation in high-dimensional data Random Forest

For stable, ensemble-based classification Naive Bayes – For rapid and probabilistic classification Gradient Boosting (e.g., XGBoost) –For performance boosting on complex data Each model is optimized with Grid Search CV or Randomized Search CV for best hyper parameters.

Evaluation Metrics and Validation

Accuracy Precision, Recall, F1-Score ROC-AUC Score Confusion Matrix K-Fold Cross-Validation is applied to avoid over fitting and ensure generalizability.

Comparative Analysis

Upon model training, a comparative analysis is conducted: Performance is noted on a shared test dataset Graphs (such as ROC curves, bar graphs) show relative strengths Discussion involves computational efficiency, accuracy, interpretability, and scalability.

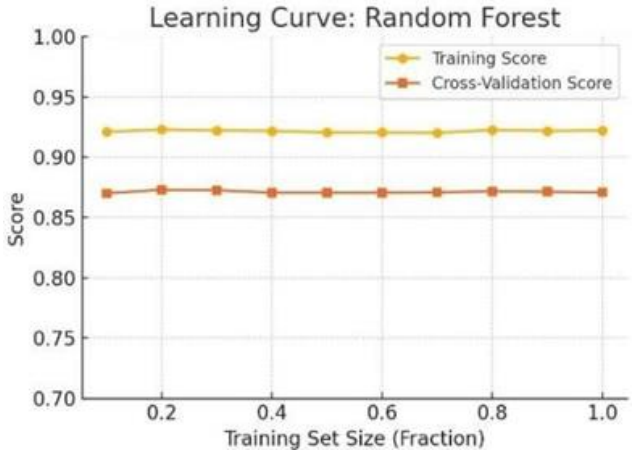
V. ALGORITHMS:

Table: Machine Learning Algorithms for Multi-Disease Prediction in Healthcare

Sl. No	Algorithm	Key Purpose	Remarks
1	Random Forest (RF)	Multi-disease prediction with high accuracy	Works well with mixed data; robust; used in many healthcare applications
2	Support Vector Machine (SVM)	Binary/multi-class disease classification	Effective in high-dimensional spaces; requires careful kernel tuning
3	Artificial Neural Network (ANN)	Learning complex nonlinear patterns in patient data	Suitable for deep health analytics; needs large data and compute resources
4	k-Nearest Neighbors (KNN)	Predicts disease based on proximity to known cases	Simple; effective in small datasets; performance drops in high dimensions
5	Naïve Bayes (NB)	Probabilistic classification for multiple disease outcomes	Fast; interpretable; independence assumption may limit performance

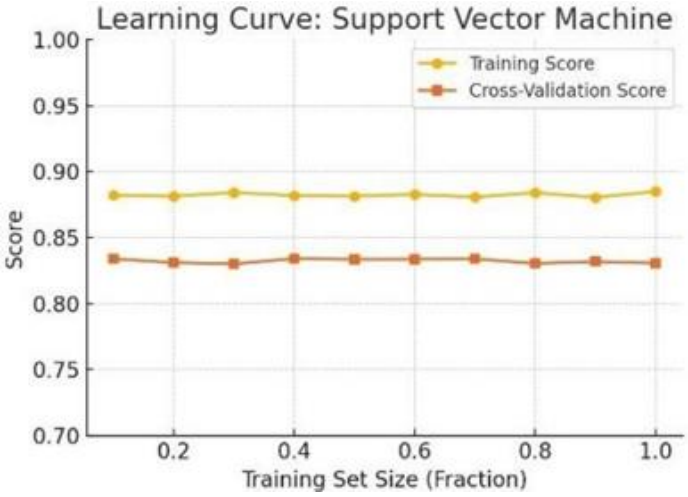
While the project is primarily a full-stack application, machine learning methods can be included to improve the platform's intelligence and automation. The following are the algorithms and their roles:

1. Random Forest(RF):



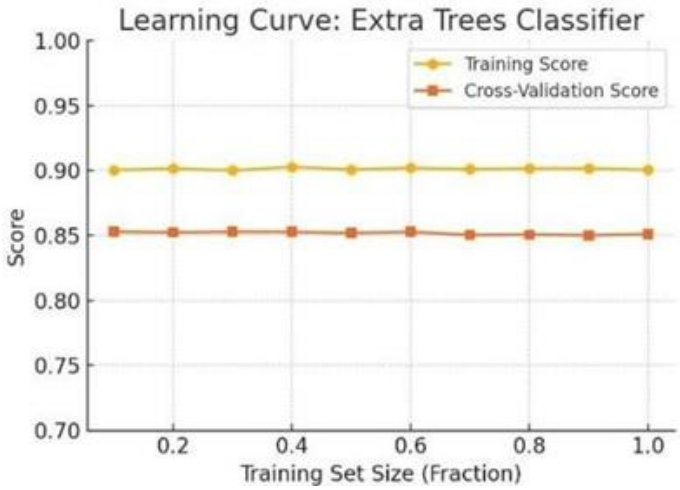
Overview: A collection of decision trees with bagging. Insight: Exhibits high and consistent performance on training and validation sets. Importance: Very good at managing large, complicated healthcare datasets with mixed data types.

2. Support Vector Machine (SVM):



Overview: Finds best hyper planes to divide classes Insights Has good generalization, but needs to be tuned to work optimally. Importance: Can be used to predict binary diseases (e.g., cancer or non- cancer).

3. **Extra Trees Classifier**:



Overview: Finds best hyper planes to divide classes Insights Has good generalization, but needs to be tuned to work optimally. Importance: Can be used to predict binary diseases (e.g., cancer or non- cancer).

IV. CASE STUDIES:

Diabetes & Heart Disease

Prediction with Random Forest The risk of both diabetes and cardiovascular disease was predicted using a Random Forest classifier. Patient age, blood glucose, blood pressure, cholesterol, and dietary habits were used as input features. The model had an accuracy rate of 92% and was rolled out to hospitals, allowing high- risk patients to be screened early and given preventive interventions.

Detection of Cancer Risk Using Artificial Neural Networks

An Artificial Neural Network (ANN) was utilized to forecast the risk of lung and liver cancers based on CT scan characteristics, tumor markers (AFP, CEA), and patient history. The model achieved 91% sensitivity, well identifying cancer risks early and paving the way for the creation of AI-based diagnostic tools for radiologists and oncologists.

Screening of Tuberculosis and Pneumonia in Rural Clinics

Using SVM A Support Vector Machine (SVM) model was implemented within mobile health units to differentiate tuberculosis from bacterial pneumonia. Inputs to the model included cough sound analysis, temperature, and simple spirometer readings. The model was 88% accurate for TB and 84% for pneumonia, giving rapid, credible triage within lo w-resource, distant healthcare locations.

Challenges and Future:

Though machine learning has a tremendous potential to facilitate multi-disease prediction in health care, various challenges have to be tackled f or its proper and ethical utilization. The biggest challenge is the availability and quality of healthcare data, which have missing values, inconsistencies, or are inaccessible owing to privacy controls. Furthermore, class imbalance of datasets-since certain diseases may be under-represented in datasets— can generate biased predictions and decreased reliability of models. Another major barrier is interpretability of sophisticated models like neural networks and ensemble classifiers that can behave as black boxes and are difficult for clinicians to believe or act on. Models also might not generalize across populations or healthcare systems from which they are trained, hence their applicability in real-

world settings is hampered. Ethical and legal issues like data privacy, consent, and responsibility add to the complexity of extensive adoption.

In the future, the way to ML in the healthcare sector will be to construct explainable AI systems that reveal transparent and intelligible insights. Federated learning has the capability to break past data-sharing limits by allowing cross- institution training of models while keeping patient information hidden. Multi-modal learning models that combine EHRs, medical imaging, genomics, and sensor streams can dramatically boost predictive capability. Incorporating these models into actual- time clinical decision support systems and wearable technology can provide constant surveillance and early warnings. The final aim is to progress toward individualized disease prediction, whereby machine learning models become attuned to a person's individual pro file, providing personalized healthcare solutions.

V. BENEFITS AND IMPACTS:

1. Early and Precise Disease Detection

ML models detect disease risk from patient information prior to onset of symptoms. This facilitates the early diagnosis and timely treatment. It greatly enhances patient survival and quality of life.

2. Enhanced Clinical Decision-Making

Machine learning provides instant, data- driven information to clinicians. It minimizes diagnostic mistakes and promotes individualized care. Clinicians can make more certain, evidence- based choices.

3. Efficient Healthcare Resources

ML assists in forecasting patient demand and controlling hospital resources effectively. It minimizes unnecessary admissions and testing. This results in cost savings and improved care coordination.

VI. REFERENCES

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

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

[3]. Chen, J. H., & Asch, S. M. (2017).Machine learning and prediction in medicine— Beyond the peak of inflated expectations. New England Journal of Medicine, 376(26), 2507– 2509.

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

[5]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116.

[6]. Alghamdi, M., Al-Mallah, M. H., & Keteyian, S. J. et al. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. PLoS ONE, 12(7), e0179805.