Disease Prediction Using Machine Learning

Mr. Ramdas Pandurang Bagawade1,
Research Scholar,

Dr. Thirupurasundari D.R.2
Associate Professor Computer Department

Bharath Institute of Higher Education and Research, India (Chennai)

Abstract: The advent of machine learning (ML) has significantly enhanced the capabilities of predictive modeling in the medical field. This study aims to evaluate the performance of various machine learning algorithms, including Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and AdaBoost, in predicting a range of diseases. Specifically, the diseases under consideration are Heart Disease, Diabetes, Kidney Disease, Parkinson's Disease, Breast Cancer, Spine Disease, and Liver Disease.

The primary objective is to assess the accuracy, computational efficiency, and interpretability of these algorithms when applied to medical datasets [1]. By leveraging large-scale health data, we develop and test predictive models to determine which algorithms offer the most reliable and efficient predictions for each disease category. The findings of this research are expected to contribute to the development of robust disease prediction models that can facilitate early diagnosis and personalized treatment plans, ultimately improving patient outcomes.

This study underscores the potential of machine learning in transforming healthcare by providing actionable insights from complex medical data. The comparative analysis of these algorithms will highlight their respective strengths and limitations, guiding future research and application in clinical settings.

Introduction

In recent years, the integration of machine learning (ML) techniques in the medical field has revolutionized the way diseases are predicted and diagnosed. Machine learning, a subset of artificial intelligence (AI), employs statistical methods to identify patterns within large datasets, enabling the development of predictive models that can assist in early disease detection and improve patient outcomes.

This paper explores the application of various machine learning algorithms, including Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and AdaBoost, in the prediction of diseases. Each of these algorithms offers unique advantages in terms of accuracy, computational efficiency, and interpretability, making them suitable for different types of medical data and prediction tasks.

The primary objective of this study is to evaluate the performance of these algorithms in predicting diseases such as Heart Disease, diabetes, Kidney Disease, Parkinson's Disease, Breast Cancer, Spine Disease and Liver Disease. By comparing the results obtained from each algorithm, we aim to identify the most effective approaches for disease prediction and highlight the potential of machine learning in enhancing clinical decision-making processes.

The significance of this research lies in its potential to contribute to the development of reliable and efficient disease prediction models. These models can facilitate early diagnosis, personalized treatment plans, and ultimately, better healthcare outcomes. As the healthcare industry continues to generate vast amounts of data, the role of machine learning in transforming this data into actionable insights becomes increasingly critical.

1) Support Vector Machine SVM [6], [9], [10], [15]

It is a supervised machine learning problem where we try to find a hyperplane that best separates the two classes. Note: Don't get confused between SVM and logistic regression. Both the algorithms try to find the best hyperplane, but the main difference is logistic regression is a probabilistic approach whereas support vector machine is based on statistical approaches. Different SVM variants, such as v-classification and one-class classification, allow for flexibility in managing class imbalance and detecting outliers. Moreover, SVMs can be enhanced through techniques like grid search for parameter optimization and k-fold cross-validation for performance evaluation

2) K-Nearest Neighbor(KNN) [4], [9], [11], [15]

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

KNN can be computationally intensive, particularly with large datasets, as it requires distance calculations for all points in the training set. To mitigate this, various strategies like locality-sensitive hashing or dimensionality reduction techniques are employed to increase computational efficiency and speed up the search for neighbors. These advancements are crucial as they enable KNN to be utilized more effectively in real-world scenarios involving vast amounts of data.

3) Logistic regression[4], [9], [10], [11], [15]

It is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Logistic regression is a statistical method widely utilized for predicting the probability of categorical response outcomes, particularly when the dependent variable is binary6. This method provides a robust approach for modeling relationships between the dependent variable and one or more independent variables.

4) Decision Tree [4], [6], [9], [10], [11], [15]

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees. Decision trees can be prone to overfitting, particularly with noisy data or when the tree becomes excessively deep. This overfitting can negatively impact model performance on unseen data. Therefore, techniques such as pruning are often employed to mitigate this issue and enhance the generalization capability of the model.

5) Random Forest [9], [10], [15]:

A **random forest** is a versatile and widely-used machine learning algorithm that combines the output of multiple decision trees to reach a single result. It can handle both classification and regression tasks. The algorithm works by creating an ensemble of decision trees, each trained on a random subset of the data and features. This approach helps improve accuracy and reduce over fitting. One of the primary advantages of Random Forest is its ability to handle missing values and maintain accuracy even when a significant proportion of the dataset is absent. However, despite its strengths, the model can be computationally intensive and may require fine-tuning of parameters to achieve optimal performance.

6) **Gradient boosting**

It is a powerful machine learning technique used for both classification and regression tasks. It works by sequentially adding models, typically decision trees, to correct the errors made by previous models. Each new model is trained to minimize the residual errors of the combined ensemble. This iterative process helps create a strong predictive model from several weaker ones, improving overall accuracy and performance. One of the primary advantages of gradient boosting is its ability to handle different types of data, including both numerical and categorical variables, without requiring extensive preprocessing. Moreover, its flexibility allows it to be utilized with various loss functions, making it adaptable to specific problem contexts.

7) AdaBoost [10]:

AdaBoost, short for Adaptive Boosting, is a machine learning algorithm designed to improve the performance of weak classifiers. It works by combining multiple weak learners, typically decision stumps, to create a strong classifier. The algorithm iteratively adjusts the weights of misclassified instances, giving more focus to difficult cases in subsequent rounds.

The effectiveness of AdaBoost lies in its ability to reduce bias and variance through the ensemble approach, making it a robust choice for many machine learning problems. Its unique adaptability to misclassification allows it to continuously improve performance over iterations, a feature that has been instrumental in various successful applications of the algorithm

Results and discussion -

All the data set used for execution of algorithms is from [17] F1 Score and Recall are generally the most critical metrics for medical diagnosis tasks, as they help balance the need to detect true cases of disease while minimizing false negatives. AUC-ROC is also valuable for understanding model performance in terms of its discriminative ability.

Accuracy:

I. When the dataset is balanced, and you want a simple measure of overall performance. If all types of errors (false positives and false negatives) have similar costs.

Limitations:

ii .In imbalanced datasets, accuracy can be misleading. For example, if 95% of patients do not have a particular disease, a model that predicts "no disease" for everyone will have high accuracy but is not useful.

Precision

i. When the cost of false positives is high. For example, in predicting a disease where a false positive could lead to unnecessary treatments or anxiety.
ii. ii. Helps assess how many of the predicted positive cases are actually positive, which is important in contexts where false positives are problematic.

Results DataFrame for Heart Disease

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.809756	0.761905	0.914286	0.831169	0.92981
Random Forest	1	1	1	1	1
SVM	0.926829	0.916667	0.942857	0.929577	0.977143
KNN	0.863415	0.873786	0.857143	0.865385	0.962905
Gradient Boosting	0.97561	0.971698	0.980952	0.976303	0.987619
Decision Tree	0.985366	1	0.971429	0.985507	0.985714
AdaBoost	0.863415	0.840708	0.904762	0.87156	0.949429

for heart random forest is giving the best results

Results DataFrame for Diabetes

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.951891	0.785714	0.590244	0.674095	0.952431
Random Forest	0.970806	0.978571	0.668293	0.794203	0.945546
SVM	0.957237	0.971963	0.507317	0.666667	0.892591
KNN	0.946957	0.839286	0.458537	0.59306	0.870555
Gradient Boosting	0.969161	0.957746	0.663415	0.783862	0.972618
Decision Tree	0.949424	0.688073	0.731707	0.70922	0.850586
AdaBoost	0.970395	0.965035	0.673171	0.793103	0.968376

For diabetes disease: Random forest is giving the best results

Results DataFrame for Kidney Disease

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.9875	0.980392	1	0.990099	0.972667
Random Forest	0.9875	0.980392	1	0.990099	0.979667
SVM	0.925	0.892857	1	0.943396	0.969333
KNN	0.9125	0.90566	0.96	0.932039	0.979667
Gradient Boosting	0.975	0.98	0.98	0.98	0.967333
Decision Tree	0.975	0.98	0.98	0.98	0.973333
AdaBoost	0.975	0.961538	1	0.980392	0.977333

For kidney Disease: Random forest is giving the best results

Results DataFrame for Parkinson's Disease

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.923077	0.933333	0.965517	0.949153	0.924138
Random Forest	0.923077	0.933333	0.965517	0.949153	0.967241
SVM	0.923077	0.90625	1	0.95082	0.955172
KNN	0.923077	0.933333	0.965517	0.949153	0.963793
Gradient Boosting	0.923077	0.964286	0.931034	0.947368	0.968966
Decision Tree	0.794872	0.888889	0.827586	0.857143	0.763793
AdaBoost	0.948718	1	0.931034	0.964286	0.986207

For Parkinson's: SVM is giving the best results

Results DataFrame for Breast Cancer

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.982456	0.96875	0.984127	0.976378	0.997942
Random Forest	0.97076	0.983333	0.936508	0.95935	0.996914
SVM	0.976608	0.968254	0.968254	0.968254	0.99662
KNN	0.959064	0.951613	0.936508	0.944	0.978689
Gradient Boosting	0.959064	0.951613	0.936508	0.944	0.99515
Decision Tree	0.935673	0.882353	0.952381	0.916031	0.939153
AdaBoost	0.976608	0.968254	0.968254	0.968254	0.996179

For Breast Cancer: Logistic Regression is giving the best results

Results DataFrame for Spine Data

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.811111	0.612903	0.791667	0.690909	0.922348
Random Forest	0.822222	0.633333	0.791667	0.703704	0.903409
SVM	0.788889	0.592593	0.666667	0.627451	0.868056
KNN	0.788889	0.586207	0.708333	0.641509	0.832386
Gradient Boosting	0.855556	0.677419	0.875	0.763636	0.91351
Decision Tree	0.8	0.607143	0.708333	0.653846	0.770833
AdaBoost	0.822222	0.642857	0.75	0.692308	0.870581

For Spine Data: Gradient Boosting is giving the best results

Results DataFrame for Liver Data

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.655172	0.654971	0.99115	0.788732	0.802989
Random Forest	0.666667	0.679739	0.920354	0.781955	0.779341
SVM	0.649425	0.649425	1	0.787456	0.704048
KNN	0.66092	0.692857	0.858407	0.766798	0.669882
Gradient Boosting	0.643678	0.675862	0.867257	0.75969	0.690846
Decision Tree	0.683908	0.719697	0.840708	0.77551	0.617075
AdaBoost	0.666667	0.706767	0.831858	0.764228	0.71754

For Liver Data: Decision Tree is giving the best results

Conclusion:

For Heart Disease random forest gives 100% accuracy which is good compared to Logistic regression, SVM, KNN, Gradient Boosting and Decision Tree. For diabetes disease random forest gives 97% accuracy which is good compared to Logistic regression, SVM, KNN, Gradient Boosting and Decision Tree. For Kidney disease random forest gives 98% accuracy which is good compared to Logistic regression, SVM, KNN, Gradient Boosting and Decision Tree. For Parkinson's Disease SVM gives 92% accuracy which is good compared to Logistic regression, random forest, KNN, Gradient Boosting and Decision Tree. For Breast Cancer Logistic regression gives 92% accuracy which is good compared to SVM, random forest, KNN, Gradient Boosting, AdaBoost and Decision Tree. For Spin data Gradient Boosting gives 85% accuracy which is good compared to Logistic regression, SVM, random forest, KNN, AdaBoost and Decision Tree. For Liver data Decision Tree gives 85% accuracy which is good compared to Logistic regression, SVM, random forest, KNN, AdaBoost and Gradient Boosting . It is observers that for liver and spine dieses the input data is contains more null values which affected the result of all algorithms.

References:

[1] Nidhi Kosarkar; Pallavi Basuri; Poonam Karamore; Prachi Gaw, "Disease Prediction using Machine Learning", International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22), 2022.

[2] Sneha Grampurohit and Chetan Sagarnal, "Disease Prediction using Machine Learning Algorithms", International Conference for Emerging Tehnology (INCET), 2020.

[3] Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel and A.D. Prajapati, "Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology, vol. 7, no. 5, pp. 2000-2002, 2020.

[4] Kedar. Pingale, Sushant. Surwase, Vaibhav. Kulkarni, Saurabh. Sarage and Abhijeet Karve, "Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology, vol. 6, no. 12, pp. 2810-2813, 2019.

[5]. "Disease Prediction in Data Mining Techniques", International Journal of Computer Applications and Information Technology (IJCAIT), 2013.

[6] "Disease Prediction System using Support Vector Machine and Multi-linear Regression", (IJIRCST), 2020.

[7] S. Jadhav, R. Kasar, N. Lade, M. Patil and S. Kolte, "Disease Prediction by Machine Learning from Healthcare Communities", International Journal of Scientific Research in Science and Technology, pp. 29-35, 2019.

[8] D. Dahiwade, G. Patle and E. Meshram, "Designing disease prediction model using machine learning approach", Proceedings of the 3rd International Conference on Computing Methodologies and Communication ICCMC 2019, no. Iccmc, pp. 1211-1215, 2019.

[9] S. Uddin, A. Khan, M. E. Hossain and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction", BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1-16, 2019.

[10] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction", 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1-4, 2018.

[11] K. Pingale, S. Surwase, V. Kulkarni, S. Sarage and A. Karve, Disease Prediction using Machine Learning, 2019.

[12] Satyabhama Balasubramanian and Balaji Subramanian, "Symptom based disease prediction in medical system by using Kmeans algorithm", International Journal of Advances in Computer Science and Technology, vol. 3.

[13]. Gaurav Shilimkar, "Disease Prediction using Machine Learning", International Journal of Scientific Research in Science and Technology, vol. 8, no. 3, pp. 551-555, 2021.

[14]. "Decision Tree based Health Prediction System", International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2020.

[15] "Disease Prediction and Hospital Recommendation using Machine Learning Algorithms", International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2020.

[16] "A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives", Turkish Journal of Computer and Mathematics Education, vol. 12, no. 6, pp. 4013-4023, 2021.

[17] https://www.kaggle.com/datasets?fileType=csv

[18] https://en.wikipedia.org/wiki/Gradient_boosting

Mr. Ramdas Pandurang Bagawade

Mr. Ramdas Pandurang Bagawade is an assistant professor in Government College of Engineering & Research, Avasari Khurd . He received his B.E. degree (2008) and M.E. degree (2013) from Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering, Pune University and he is pursuing Phd at Bharath Institute of Higher Education and Research, Chennai.

Dr. Thirupurasundari D R

Dr. Thirupurasundari is an associate professor in Bharath Institute of Higher Education and Research, Chennai. she received her M.E. degree (2011) from Anna University and Phd degree (2022) from Vyankateshwara University, Gaziyabad. Her area of research include Network Security, Machine Learning and Wireless Communication.