Precision in Classification : A Comparative Study of SVM, KNN, Random Forest, Naïve Bayes for Early Disease Detection

MVSP Praneeth (21B81A05X5, CSE E)

Abstract: This article dives deep into the current state of healthcare prediction, offering a thorough and insightful analysis. It highlights the tremendous benefits that have come from integrating artificial intelligence, emphasizing the positive impact it has had. While the use of AI in healthcare prediction has brought about significant advancements, it also presents its own set of challenges. The goal of this article is to contribute to the progress of disease detection and prediction by presenting the findings of a comprehensive review of recent research articles in the field. Additionally, it explores the potential impact of these findings. Healthcare prediction has become vital in saving lives, and intelligent systems have emerged to analyse complex data relationships and generate valuable insights for predictions. The paper carefully examined numerous working papers, shedding light on the methodologies employed in each study. It also acknowledges the challenges that need to be addressed in order to fully harness the potential of artificial intelligence in disease diagnosis and prediction, and proposes potential solutions to overcome these challenges. Research has shown that AI plays a significant role in accurately diagnosing diseases, anticipating healthcare needs, and analysing health data. By leveraging vast amounts of clinical records and reconstructing patients' medical histories, AI has proven to be a powerful tool in the healthcare sector.

1. Introduction

In today's world, people come across a wide range of diseases due to their current environment and lifestyle choices. It's crucial to identify and predict these illnesses early on to prevent them from becoming severe. According to medical reports, chronic diseases contribute to an increase in human mortality rates. Some common chronic illnesses include diabetes, cardiovascular diseases, cancer, strokes, hepatitis C, and arthritis. Given their long-lasting nature and significant impact on mortality, accurately diagnosing these conditions is extremely important in the healthcare sector. So, it's crucial to address the factors that contribute to a patient's risk of mortality.

Advancements in medical research have made it easier to collect health-related data. Machine learning can help analyze patient data and other relevant information, making early disease detection possible. In the field of machine learning, there are various techniques available, such as semi-supervised learning, supervised learning, unsupervised learning, and deep learning.

To meet this need, it's essential to develop a machine learning model that can take input symptoms and predict the probability and risk of disease progression, as well as its impact on an individual's well-being. The main goal is to use a machine learning approach to identify and predict chronic diseases in individuals. The dataset, which is a crucial part of this process, consists of two types of information. First, there is structured data that includes details like the patient's age, gender, height, weight, and more . However, this structured data intentionally excludes any personal identifiers, such as the patient's name or ID.

1. Technologies Used

1.1 Machine Learning

Machine learning is a subset of artificial intelligence(AI) that entails the process of training algorithms using data, enabling them to make predictions or execute actions without the need for explicitprogramming. Machine Learning involves many algorithms which comes under 2 types i.e. Supervised Learning and Unsupervised Learning. Supervised machine Learning is of two types: a) Classification b) Regression

1.1.1 Supervised learning classification algorithms:

The types of classification Supervised learning algorithms are decision trees, support vector machines, naïve Bayes, K-nearest neighbours, and neural networks.

- **Decision Tree**

The Decision Tree is used for each internal node to represent a feature tree, a leaf node to represent a class label, and branches to indicate conjunctions of features. In a non-parametric supervised learning approach, decision trees are used in both regressionand classification. Calculation of the decision tree is done using information gain and entropy. Equation 1 provides the equation for calculating the entropy:

Gain(T,X) = Entropy(T)-Entropy(T,X) (2)

Where T is the target value and X is the actual variable of the dataset [33]

which disease is positive and which is not can be categorized through this Decision tree algorithm.

- **SVM**

It combines supervised regression and classificationlearning methods. As a result, a higher dividing hyperplane can be constructed by projecting the input vector to a higher-dimensional space. [34]. It

utilizes the dividing or separating hyperplane with the expression to view this training information:

$$w.m + b=0 \quad (3)$$

In this equation 3, b is a scalar, and m is a dimensional vector, w is perpendicular to the horizontal

dividing hyperplane. As seen in Figure 2,an SVM trained on instances from 2 has maximum edge hyperplane

classes.

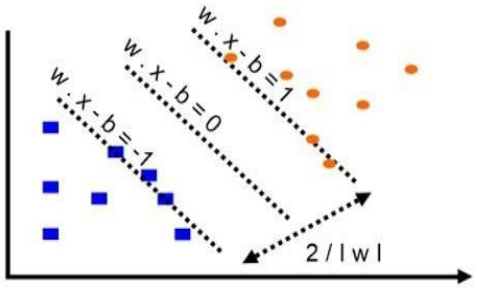

Fig 2. An SVM trained with samples from 2classes.[35]

- **NAIVE BAYES**

Naive Bayes classifiers encompass a set of classification techniques rooted in Bayes' Theorem. They constitute a family of algorithms characterized by a common principle: the independence assumption, meaning that every pair of features being classified is treated as independent of each other

- **KNN**

The K-NN operates on the assumption of similarity between the new data point and existing cases, assigning the new data point to the category that bears the closest resemblance among the available cayegories.

- **RANDOM FOREST**

The Random Forest (RF) classifier, the better it is atbeing accurate. It makes a bunch of trees called a forest that work together to predict things better. Every decision tree in the random forest is made using only part of the data and trained using estimates. The RF algorithm tries to make the best choice by putting together the results from many DTs.

With the help of this algorithm, results can be obtained in the form of tree and disease prediction can be done in a better way.

- **Feature Extraction**

Feature engineering is the practice of transforming a dataset to enhance the performance of a machine learning model during training. It is Modifying the dataset through actions such as adding, removing, merging, or altering features. This meticulous adjustment of the training data is done with the aimof ensuring that the resultant machine learning model is well-suited to meet its objectives. A variety of techniques exist for feature extraction, such as principal component analysis (PCA), autoencoders, filter methods, wrapper methods etc.

Following Table1 shows the work done on disease detection and Prediction

Results

- **Evaluation Matrix of Supervised Classification Algorithms**

The evaluation of supervised classification algorithms often involves assessing their performance using metrics such as accuracy, sensitivity, and specificity. These metrics provide insights into how well the model is performing in different aspects of classification.[39]

- Accuracy: Accuracy is a metric used to gauge the overall correctness of a model's predictions. To calculate accuracy, we divide the sum of correctly predicted instances, which includes both true positives and true negatives, by the total number of instances in the dataset.
- Sensitivity (True Positive Rate or Recall): Sensitivity gauges the model's ability to correctly identify positive instances. It is calculated as the

ratio of true positives to the total number of actual positive instances. The formula for sensitivity is:

Sensitivity = TP / (FN + TP)

- Specificity (True Negative Rate): Specificity assesses the model's capability to correctly identify negative instances. It is determined by the ratio of true negatives to the total number of actual negative instances. Specificity can be expressed as:

Specificity = TN / (FP + TN)

- F1 Score: It is calculated as:
F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
- **Evaluation Matrix of Supervised Regression Algorithms:**

Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are frequently employed metrics for assessing the effectiveness of regression models.

Conclusion

In conclusion, predictive health refers to the use of predictive analytics and ML and DL algorithms to improve health and healthcare services. The adoption of machine learning and deep learning techniques has the potential to revolutionize traditional healthcare delivery. Healthcare data is recognized as a crucial component that contributes to the advancement of medical-care systems. The availability of diverse sources of health data has increased tremendously in current years.

The history of a predictive analytics tool, its field of use, and its approach for predicting Disease have allbeen covered in this paper.

The paper discussed the background of a Predictive Analytics Tool and its domain, focusing on the methodology for early disease prediction. It emphasizes the potential of artificial intelligence (AI) in enhancing the quality of work in healthcare. The paper reviewed many working papers and provided insights into the methodologies employed in each study. This paper also finds out limitations of studied research paper and suggests possible solution for it. Research has demonstrated that AI plays a significant role in accurate disease diagnosis, healthcare anticipation, and analysis of health data by leveraging large-scale clinical records and reconstructing patients' medical histories. However, further studies are needed to improve AI integration with healthcare data quality management considerations.