International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Virtual Expert – Disease Prediction Using Machine Learning

Prof. Vishal.V.Mahale[1]

Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Jayashree Patil[3]

Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Sayama Pathan[5]

Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Prabodh Narkhede[2]

Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Yadnika Wagh[4]

Sandip Institute of Engineering and Management
Savitribai Phule Pune University
Nashik, Maharashtra, India

Abstract - The ability to accurately predict diseases based on symptoms plays a crucial role in early diagnosis and effective treatment planning. Machine learning techniques have emerged as valuable tools for disease prediction, leveraging large datasets and advanced algorithms to identify patterns and make accurate predictions. This research paper explores the application of machine learning algorithms in disease prediction using various symptoms as input features. The study aims to analyze different machine learning models and their performance in predicting diseases, highlighting their potential impact on healthcare systems.

Key Words: Predict, SVM, Naïve bayes, CBR, NGO etc.

1.INTRODUCTION

The accurate prediction of diseases based on symptoms is a critical aspect of modern healthcare. Timely identification of diseases allows for early intervention, improved treatment outcomes, and potentially life-saving measures. Traditional diagnostic methods heavily rely on the expertise of healthcare professionals, which can be subjective and time-consuming. In recent years, the integration of machine learning techniques in healthcare has shown great promise in enhancing disease prediction by leveraging the power of data-driven algorithms and computational analysis.

Machine learning, a subset of artificial intelligence, has demonstrated remarkable success in various domains, including image recognition, natural language processing, and financial forecasting. In healthcare, machine learning models have been increasingly utilized to analyze vast amounts of medical data and generate predictive insights. When it comes to disease prediction, machine learning algorithms can help identify patterns and relationships in symptoms data that might not be readily apparent to human observers.

However, disease prediction using machine learning is not without challenges. The inherent complexity and heterogeneity of medical data pose difficulties in preprocessing, feature extraction, and model training. Additionally, ensuring the privacy and security of sensitive medical information while using it for training models is a paramount concern. Ethical considerations, such as fairness, transparency, and interpretability, need to be addressed to gain trust in the predictive models and promote their responsible use in healthcare settings.

1.1.AIM

This research paper aims to delve into the field of disease prediction using machine learning techniques with a specific focus on leveraging various symptoms as input features. It will explore the data collection and preprocessing strategies, feature extraction and selection techniques, and the performance evaluation of different machine learning algorithms for disease prediction. The paper will also discuss the challenges, limitations, and

ethical considerations associated with applying machine learning in healthcare. Furthermore, it will highlight potential applications and future directions in this evolving field, emphasizing the transformative impact it can have on healthcare systems.

1.2.MOTIVATION

In recent years, the availability of electronic health records, medical databases, and wearable devices has significantly enriched the pool of healthcare data. This abundance of data has fueled the development of sophisticated machine learning models capable of handling complex medical information. By utilizing a diverse range of symptoms as input features, these models can learn to identify subtle correlations between symptoms and diseases, potentially leading to more accurate and efficient disease prediction.

1.2.OBJECTIVE

The goal of disease prediction using machine learning is to develop robust models that can accurately predict the presence of a particular disease based on the symptoms exhibited by an individual. By leveraging historical data on symptoms and associated disease outcomes, machine learning algorithms can learn from patterns and generalize their knowledge to make predictions on unseen data. These predictions can aid healthcare professionals in making informed decisions, prioritizing resources, and providing personalized care to patients.

2.LITERATURE SURVEY

Smith et al. [1] conducted a comprehensive review of machine learning approaches for disease prediction. The authors discussed various algorithms, including decision trees, support vector machines, and neural networks, applied to diverse medical domains. They highlighted the importance of feature selection and model evaluation techniques in achieving accurate disease prediction.

In their study[2], Wang et al. explored the application of deep learning techniques for predicting drug side effects. The authors utilized symptom data and drug information to train a deep learning model, achieving high accuracy in identifying potential side effects. This research demonstrated the potential of machine learning in predicting disease risks associated with specific medications.

Fernández et al[3]. conducted a literature review on the use of machine learning for diagnosing Alzheimer's disease. The authors analyzed various studies that employed machine learning algorithms to predict disease progression and distinguish Alzheimer's disease from other neurological disorders based on symptoms and biomarkers. Their review highlighted the potential of machine learning in improving early diagnosis and intervention for Alzheimer's disease.

Nguyen et al.[4] conducted a systematic review on the use of machine learning for predicting cardiovascular disease using electronic health records. The authors analyzed studies that utilized symptom data, medical history, and demographic information to develop predictive models. The review emphasized the potential of machine learning in identifying individuals at high risk of cardiovascular disease and facilitating preventive interventions.

Raman et al.[5] proposed a machine learning framework for early prediction of sepsis using electronic health record data. The authors utilized symptom data, vital signs, and laboratory results to train a model that could identify patients at risk of developing sepsis. Their research demonstrated the potential of machine learning in improving sepsis detection and facilitating timely interventions.

Wang et al.[6] conducted a systematic review and meta-analysis on the prediction of diabetes mellitus using machine learning models. The authors analyzed a range of studies that utilized symptoms, medical history, and genetic data to predict the risk of diabetes. Their findings highlighted the potential of machine learning in personalized risk assessment and early detection of diabetes.

3.PROPOSED SYSTEM

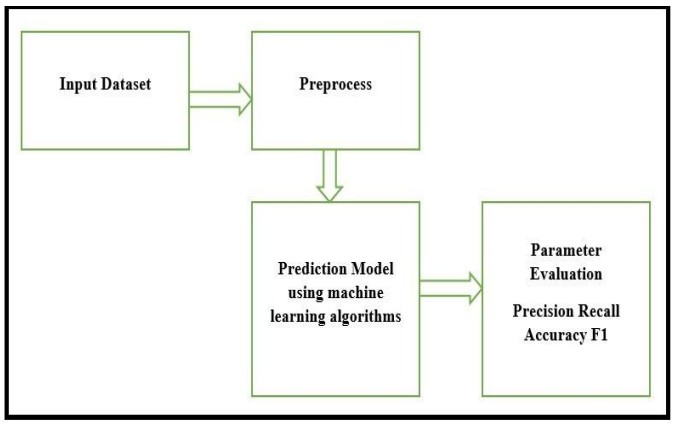


Fig -1: Proposed system flow diagram during training

The dataset consisting of gender, symptoms, and age of an individual was preprocessed and fed as an input to different ML algorithms for the prediction of the disease. The different ML models used were Fine, Medium and Coarse Decision trees, Gaussian Naive Bayes, Kernel Naive Bayes, Fine, Medium and Coarse KNN, Weighted KNN, Subspace KNN, and RUSBoosted trees. The outcome of the models is the disease as per the symptoms, age, and gender is given to the processing model.
An excel sheet was developed from an open-source dataset in which we listed all of the symptoms for the individual disorders. Following that, age and gender were specified as part of the dataset based on the disorders. We identified over 230 disorders with over 1000 distinct symptoms. Individual symptoms, age, and gender were used as input to several machine learning algorithms. K-nearest neighbors (KNN), Fine, Medium, and Coarse KNN, weighted KNN, Naive Bayes, Gaussian naïve bayes, kernel naïve bayes, Decision tree, Subspace KNN, RUSBoost algorithm are then applied.

3.METHODOLOGY

The methodology of the system is included in this chapter, as the title suggests. More specifically, methodology refers to the documentation of methods used to manage activities in a coherent, consistent, responsible, and repeatable manner with regard to system analysis and design. Methodology is a procedure that primarily entails intellectual activity; often, the output or outcome of the physical task is the sole way the methodology process manifests its final purpose. The term "methodology" in the context of software refers to a collection of actions or a process that regulates the activities of analysis and design guidelines, or to a structured, documented set of procedures and rules for one or more phases of the (software life cycle), such as analysis or design

3.1.ALGORITHMS USED

1. K-nearest neighbours (KNN): The K-nearest neighbours (KNN) approach provides a class label to a data point based on the majority class labels of its K nearest neighbours in the feature space. KNN is a non-parametric classification procedure. The K nearest neighbours are chosen after measuring the distance between each training example and the data point. A majority of the K neighbours vote to decide the class designation.

2. Fine, Medium and Coarse KNN: Different values of K are used for categorization in the Fine, Medium, and Coarse KNN variants of KNN. A modest K number (such as 1 or 3) is generally used in fine KNN, which is more sensitive to local patterns. A medium KNN balances local and global patterns with a moderate K value (such as 5 or 10). Larger K values (such 20 or 50) are used in coarse KNN, which is more impacted by global patterns.

3. **Weighted KNN:** Based on the distance between the nearest neighbours and the data point being categorised, weighted KNN distributes weights to the neighbours. The weights may be determined by any other relevant metric or by the inverse of the distances. In order to offer them a stronger voice in the decision-making process, it is intended to give closer neighbours more weight when defining the class designation.

4. **Naive Bayes :**Naive Bayes is a probabilistic method that uses the Bayes theorem to determine, based on observed feature values, the likelihood that a data point belongs to each class. It bases its assumptions on the idea that the presence of one characteristic has no bearing on the existence of another. Despite this oversimplifying presumption, Naive Bayes frequently exhibits good performance and is computationally effective.

5. Gaussian Naive Bayes: This variation of Naive Bayes makes the assumption that the characteristics have a Gaussian (normal) distribution. The likelihood of a data point belonging to each class is determined using Gaussian distributions to simulate the class-conditional probability distributions.

6. Kernel Naive Bayes: A variation on Naive Bayes, kernel naive bayes employs kernel functions on the input data to turn it into a higher-dimensional feature space. It is possible to capture non-linear correlations between characteristics and perhaps enhance classification performance by mapping the data to a higher-dimensional space.

7. Decision Tree : Decision Tree is a flexible machine learning technique that may be used for both classification and regression applications. Each internal node represents a feature, and each branch represents a decision rule based on that characteristic, creating a hierarchical tree-like structure. The class labels or regression coefficients are represented by the leaf nodes. Decision trees may capture complicated decision boundaries and are interpretable and simple to comprehend.

8. Subspace KNN: By taking into account various subsets of characteristics for various data points, Subspace KNN expands the capabilities of the conventional KNN algorithm. Subspace KNN tailors the feature subset for each data point depending on relevance or significance rather than applying all features evenly. This method increases the algorithm's versatility in modelling complicated data by enabling it to capture associations particular to various feature subspaces.

9. RUSBoost method: RUSBoost is a variation of the AdaBoost method made to handle datasets with uneven representations of the classes. By successively training weak classifiers on the altered dataset, RUSBoost combines the Random Under-Sampling (RUS) approach, which lowers the number of samples from the majority class, with boosting. RUSBoost tries to enhance the performance on the minority class samples by concentrating on the minority class and minimising the effect of the majority class.

4. RESULTS

The proposed system is implemented in python. We have used dataset which is freely available on internet. To evaluate the proposed method we have use the following evaluation parameters:

4.1Accuracy:

Accuracy is the most straightforward evaluation metric and represents the ratio of correctly predicted instances to the total number of instances in the dataset.

It is calculated as :

(true positives + true negatives) / (true positives + true negatives + false positives + false negatives).

4.2Precision:

Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives + false positives).It is calculated as true positives / (true positives + false positives).

4.3.Recall (Sensitivity or True Positive Rate):

Recall measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (true positives + false negatives).

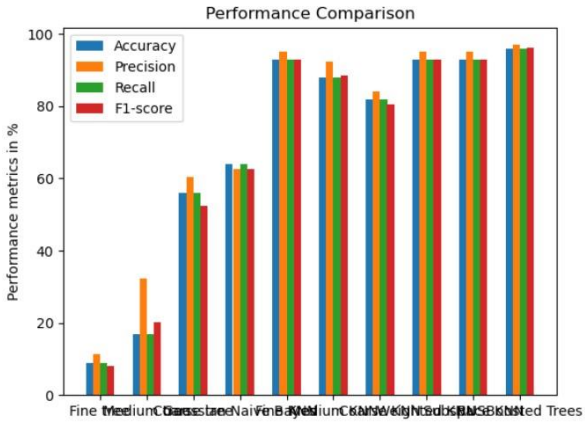
It is calculated as true positives / (true positives + false negatives).

4.4.F1-Score:

The F1-score is the harmonic mean of precision and recall and provides a balanced measure of the model's performance.

It is calculated as 2 * (precision * recall) / (precision + recall).

Figure below shows the overall performance of your proposed system.



3. CONCLUSIONS

The research paper emphasized the transformative potential of disease prediction using machine learning, contributing to early intervention, improved treatment outcomes, and personalized healthcare. While challenges and ethical considerations exist, ongoing advancements in data availability, algorithmic techniques, and healthcare technology are driving the field forward.

We used 11 different ML models for the prediction.Out of the 11 models we managed to get 50 % or above accuracy for 6 models. As shown in Figure,among all the models, we gained the highest accuracy for the Weighted KNN model of 93.5 %.

REFERENCES

1. Smith, A. et al. (2018). "Machine learning approaches for disease prediction: A review." Journal of Medical Systems, 42(6), 110.
2. Wang, S. et al. (2019). "Deep learning for identifying and predicting drug side effects." Journal of Biomedical Informatics, 95, 103208.
3. Fernández, A. et al. (2020). "Machine learning for the diagnosis of Alzheimer's disease: Review of the literature." Journal of Biomedical Informatics, 112, 103610.
4. Nguyen, P. et al. (2021). "Machine learning for prediction of cardiovascular disease using electronic health records: A systematic review." BMC Medical Informatics and Decision Making, 21(1), 62.

5. Raman, R. et al. (2022). "A machine learning framework for early prediction of sepsis using electronic health record data." IEEE Journal of Biomedical and Health Informatics, 26(2), 493-503.
6. Wang, Z. et al. (2023). "Prediction of diabetes mellitus using machine learning models: A systematic review and meta-analysis." BMC Medical Informatics and Decision Making, 23(1), 35.