**Multiple Disease Prediction**

PROJECT SYNOPSIS OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

Computer Science and engineering

SUBMITTED BY

**VIRAJ SINGH**

**SUDHEER KUMAR SINGH**

**VARUN SAINI**

Under the supervision of

**(Prof. Dharmendra Kumar)**



Department of Computer Science and Engineering

**KIET Group of Institutions, Delhi-NCR,**

**Ghaziabad (UP)**

|  |  |
| --- | --- |
| **Content** | **Page no.** |
| Introduction | 3 |
| Aim | 4 |
| Need | 5 |
| Data Section | 6 |
| Building Model | 7 |
| Literature Review | 8 |
| Conclusion | 13 |
| References | 14 |
|  |  |

**Table of content**

**INTRODUCTION**

The wide adaptation of computer-based technology in the health care industry resulted in the accumulation of electronic data. Due to the substantial amounts of data, medical doctors are facing challenges to analyze symptoms accurately and identify diseases at an early stage. However, supervised machine learning (ML) algorithms have showcased significant potential in surpassing standard systems for disease diagnosis and aiding medical experts in the early detection of high-risk diseases. In this literature, the aim is to recognize trends across various types of supervised ML models in disease detection through the examination of performance metrics. The most prominently discussed supervised ML algorithms were Naïve Bayes (NB), Decision Trees (DT), K-Nearest Neighbour (KNN). As per findings, Support Vector Machine (SVM) is the most adequate at detecting kidney diseases and Parkinson's disease. The Logistic Regression (LR) performed highly at the prediction of heart diseases. Finally, Random Forest (RF), and Convolutional Neural Networks (CNN) predicted in precision breast diseases and common diseases, respectively.

**AIM**

The aim of this project is to test the proposed hypothesis that supervised ML algorithms can improve health care by the accurate and early detection of diseases. In this study, we investigate studies that utilize more than one supervised ML model for each disease recognition problem. This approach renders more comprehensiveness and precision because the evaluation of the performance of a single algorithm over various study settings induces bias which generates imprecise results. The analysis of ML models will be conducted on few diseases located at heart, kidney, breast, and brain. For the detection of the disease, numerous methodologies will be evaluated such as KNN, NB, DT, CNN, SVM, and LR.

**NEED**

In recent years, the healthcare domain is evolving more due to the integration of information technology (IT) in it. The intention to integrate IT in healthcare is to make the life of an individual more affordable with comfort as smartphones made one’s life easier [3]. This could be possible by making healthcare to be intelligent, for instance, the invention of the smart ambulance, smart hospital facilities, and so on, which helps the patients and doctors in several ways [4]. The research on a specified region for patients affected by chronic diseases every year had been held and found that the difference between the patients in gender wise is very small, and it is found that the large number of patients were admitted in the year 2014 for treating chronic diseases. The use of structured and unstructured data provides highly accurate results instead of using only structured data. Since the unstructured data includes the doctor’s records on the patients related to diseases and the patient’s symptoms and grievances faced by them, explained by themselves, which is an added advantage when used along with the structured data that consists of the patient demographics, disease details, living habitats, and laboratory test results [5, 6]. It is difficult to diagnose rare diseases. Hence, the use of self-reported behavioural data helps differentiate the individuals with rare diseases from the ones with common chronic diseases. By using machine learning approaches along with questionnaires, it is believed that the identification of rare diseases is highly possible.

**SECTIONS**

**Data collection**

Data collection has been done from the internet to identify the disease here the real symptoms of the disease are collected i.e. no dummy values are entered. The symptoms of the disease are collected from different health related websites. Data Pre-processing Before feeding the data into the Prediction model, following data cleaning and pre-processing steps are performed

● Checking null values and filling using forward fill method

● Converting data into different cases

● Standardizing the data using mean and standard deviation

● Splitting the dataset into training and testing sets.

**Building Model**

Many methods are used to perform data mining. Machine learning is one of the approaches. Random forest Machine learning strategies include grouping, clustering, summarization, and many others. Since classification techniques are used in this project, classification is one of the data mining processes in this phase of categorical data classification. And this step is divided into two phases: training and testing. In the training phase, predetermined data and associated class labels are used for classification. The training stage is often referred to as supervised learning. The preparation and testing phases of the classification process are depicted in the diagram. In the training process, training tuples are used, and in the test data phase, test data tuples are used, and the classification rule's accuracy is calculated. Assume that the classification rule's accuracy on testing data is sufficient for the rule to be used for classification of unmined data.

LITERATURE REVIEW

**A.** **Diabetes Diseases: -**

The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques.

Dr Saravana Kumar N M, Eswari, Sampath P and Lavanya S (2015) implemented a system using Hadoop and Map Reduce technique for analysis of Diabetic data. This system predicts type of diabetes and also risks associated with it. The system is Hadoop based and is economical for any healthcare organization. Aiswarya Iyer (2015) used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result. K. Rajesh and V. Sangeetha (2012) used classification technique.

They used C4.5 decision tree algorithm to find hidden patterns from the dataset for classifying efficiently. Humar Kahramanli and Novruz Allahverdi (2008) used Artificial neural network (ANN) in combination with fuzzy logic to predict diabetes. B.M. Patil, R.C. Joshi and Durga Toshniwal (2010) proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm. In order to build classifiers C4.5 decision tree algorithm is used.Mani Butwall and Shraddha Kumar (2015) proposed a model using Random Forest Classifier to forecast diabetes behaviour.Nawaz Mohamudally1 and Dost Muhammad (2011) used C4.5 decision tree algorithm, Neural Network, K-means clustering algorithm and Visualization to predict diabetes.

B. Heart Diseases:-

Marimuthu aimed to predict heart diseases using supervised ML techniques. The authors structured the attributes of data as gender, age, chest pain, gender, target and slope. The applied ML algorithms that were deployed are DT, KNN, LR and NB. As per analysis, the LR algorithm gave a high accuracy of 86.89%, which deemed to be the most effective compared to the other mentioned algorithms.

In 2018, Dwivedi attempted to add more precision to the prediction of heart diseases by accounting for additional parameters such as Resting blood pressure, Serum Cholesterol in mg/dl, and Maximum Heart Rate achieved. The used dataset was imported from the UCI ML laboratory; it was comprised with 120 samples that were heart disease positive, and 150 samples that were heart disease negative. Dwivedi attempted to evaluate the performance of Artificial Neural Networks (ANN), SVM, KNN, NB, LR and Classification Tree.

At the appliance of tenfold cross validation, the results showed that LR has the highest classification accuracy and sensitivity, which shows high dependability at detecting heart diseases. This conclusion is strengthened by the findings of Polaraju and Vahid et al. Where the Logistic Regression outperformed other techniques such as ANN, SVM, and Adaboost. The studies excelled in conducting an extensive analysis on the ML models.

For instance, various hyper-parameters were tested at each ML algorithm to converge to the best possible accuracy and precision values. Despite that advantage, the small size of the imported datasets constraints the learning models from targeting diseases with higher accuracy and precision.

C. Parkinson’s Disease: -

Chen et al presented an effective diagnosis system using Fuzzy k-Nearest Neighbor (FKNN) for the diagnosis of Parkinson’s disease (PD) The study focused on comparing the proposed SVM-based and the FKNN-based approaches. the Principal Component Analysis (PCA) was utilized to assemble the most discriminated features for the construction of an optimal FKNN model.

The dataset was taken from the UCI depository, and it recorded numerous biomedical voice measurement ranging from 31 people, 24 with PD. The experimental findings have indicated that the FKNN approach advantageously achieves over the SVM methodology in terms of sensitivity, accuracy, and specificity. In line of this study, Behroozi aimed to propose a new classification framework to diagnose PD, which was enhanced by a filter-based feature selection algorithm that increased the classification accuracy up to 15%.

The classification of the framework was characterized by applying independent classifiers for each subset of the dataset to account for the loss of valuable information. The chosen classifiers were KNN, SVM, Discriminant Analysis and NB. The results showed that SVM achieved the highest in all the performance metrics. In addition, Eskidere concentrated on tracking the progression of PD by discussing the performance of SVM with other classifiers such as Least Square Support Vector (LS-SVM), General Regression Neural Network (GRNN) and Multi-layer Perceptron Neural Network (MLPNN). The findings indicated that LS-SVM is the highest performing model. This conclusion is strengthened by the adequate comparison of decoders with their optimal performance metric.

According to Lavesson , various ML algorithms are designed to optimize numerous performance metrics (e.g., Neural Networks optimizes squared error whereas KNN and SVM optimize accuracy). Furthermore, the authors are particularly good at proposing frameworks with details. For example, SVMs parameters such as the kernel

**CONCLUSION**

The use of different ML algorithms enabled the early detection of many maladies such as heart, kidney, breast, and brain diseases. Throughout the literature, SVM, RF and LR algorithms were the most widely used at prediction, while accuracy was the most used performance metric. The CNN model proved to be the most adequate at predicting common diseases. Furthermore, SVM model showed superiority in accuracy at most times for kidney diseases and PD because of its reliability in handling high-dimensional, semi-structured and unstructured data. For Breast cancer prediction, RF showed more superiority in the probability of correct classification of the diseases because of its ability to scale well for large datasets and its susceptibility to avoid overfitting. Finally, the LR algorithm proved to be the most reliable in predicting heart diseases. In future work, the creation of more complex ML algorithms is much needed to increase the efficiency of disease prediction. In addition, learning models should be calibrated more often after the training phase for potentially a better performance. Moreover, datasets should be expanded on different demographics to avoid overfitting and increase the accuracy of the deployed models. Finally, more relevant feature selection methods should be used to enhance the performance of the learning models.

**References: -**

[1] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, “Prediction of heart disease using machine learning,” in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.

[2] Y. Hasija, N. Garg, and S. Sourav, “Automated detection of dermatological disorders through image-processing and machine learning,” in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 1047–1051.

[3] 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.

[4] R. Katarya and P. Srinivas, “Predicting heart disease at early stages using machine learning: A survey,” in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302–305.

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

[6] M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra, “A proposed model for lifestyle disease prediction using support vector machine,” in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1–6.