Multiple Disease Detection System using Machine Learning

Pooja Vajpayee [1], Aniket Raj [2], Ankit Pal [2], Ashutosh Shukla [2]

[1]*Assistant Professor, Department of Computer Science & Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, UP, India.*

[2]*Student, Department of Computer Science & Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, UP, India.*

Corresponding Author: aniketraj992@gmail.com

Abstract: - In this world man suffers from many kinds of diseases. Illnesses can affect a person physically as well as mentally. Diseases arise mainly due to four reasons: (i) infection, (ii) deficiency, (iii) heredity and (iv) dysfunction of body organs. In our company, doctors or health professionals are required to detect and diagnose the relevant disease and provide medical therapy or treatment to cure or control the disease. Some diseases are cured after treatment, but chronic diseases are never cured despite treatment; Treatment can prevent chronic conditions from getting worse over time. Therefore, it is always important to detect and treat the disease at an early stage. To help doctors or health professionals, this chapter proposes a disease detection system that doctors or health professionals can use to detect diseases in patients using the graphical user interface of DDS. DDS has been developed to detect certain diseases such as liver disorders, hepatitis, heart disease, diabetes and chronic kidney disease. Patients with each disease have different signs and symptoms. To implement DDS, various datasets are obtained from the Kaggle machine learning database. The Adaboost classifier algorithm is used for disease detection to calculate the classifier in DDS. It is a machine learning algorithm that results in the identification of diseases listed in the DDS with 100% accuracy, precision and recall. DDS GUI was built with Python support as a screening tool so that doctors or health professionals can easily find patients with this disease.

Key Words: - ***Diseases, Health professionals, Machine learning algorithm.***

I. Introduction

India is not a country where majority of the population depends on doctors. Recently, machine learning algorithms have been widely used to predict various properties, and these algorithms are widely used in various fields. Machine learning in healthcare is one of the main research areas where machine learning models are applied to medical databases to predict various behaviors. This paper compares classical machine learning models with different ensembles used to predict diabetes and heart disease risk from different data sets, and comparative analysis shows that the superlearner model provides the best accuracy of 86% for this data. The formula provides 97% accuracy for predicting diabetes and diabetes risk. In this example, we focus on developing machine learning-based algorithms for the detection and diagnosis of various diseases such as heart disease, diabetes, and Parkinson's disease. Developing a medical diagnostic system based on machine learning (ML) algorithms to predict any disease can provide a more accurate diagnosis than conventional methods. Based on symptoms, age, and gender, the diagnostic system predicts a person's likelihood of having the disease. The weighted algorithm gives the best results compared to other algorithms. One of the most important topics in society is human health. They are looking for the best and most reliable diagnosis of the disease, so they can get the treatment they need as soon as possible. Additional courses such as statistics and computer science are required for the health aspects of research as these concepts are often complex. The challenge to promote new approaches is to challenge these rules and move beyond conventional boundaries. The abundance of new methods allows for a comprehensive review that avoids some aspects. To this end, we propose a systematic analysis of human

Manuscript revised May 21, 2023; accepted May 22, 2023. Date of publication May 25, 2023.

This paper available online at www.ijprse.com

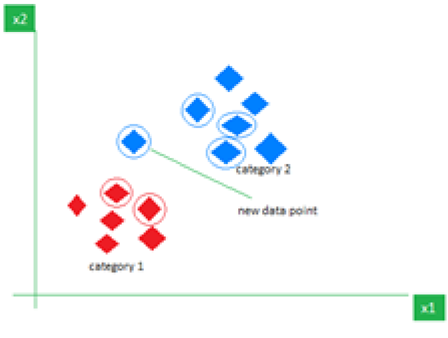
ISSN (Online): 2582-7898; SJIF: 5.59

diseases involving machine learning. With a busy and stressful lifestyle, heart disease is now on the rise. Heart disease affects all age groups, so early detection of heart disease through symptoms or reports is very important.

II. Methodology

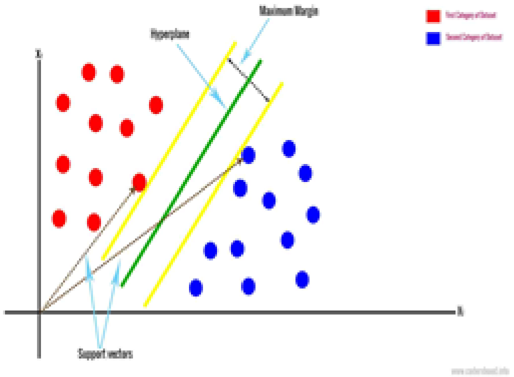
2.1 K-Nearest Neighbour Algorithm

The k-Nearest Neighbors (KNN) calculation may be a directed machine learning calculation that's broadly utilized for classification purposes. Broadly utilized for malady forecast 1. KNN may be a administered calculation that gauges the conveyance of unlabeled information, taking under consideration the characteristics and labels of the information. In common, the KNN calculation can classify information utilizing the same preparing demonstrate as the inquiry, taking into consideration the k closest information focuses (neighbors) closest to the message. At last, the calculation employments a larger part vote to decide which classification is final. Among all machine learning calculations, the KNN calculation is one of the only and most broadly utilized classification errands due to its adaptable and easy-to-understand design3. The calculation is known for utilizing information of diverse sizes, number of letters, noise level, variety and substance to solve the issue of competition and classification 4. Yes, so this article rotates around this prepare based on classification of restorative information., since anticipating infection may be a genuine challenge within the world. To unravel this issue, it is essential to think almost how it can be changed.



2.2 Support Vector Machine Algorithm

Support Vector Machine or SVM is one of the most popular supervised learning algorithms used for both classification and regression problems. However, it is primarily used for classification problems in machine learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate the n-dimensional space into classes so that we can easily place a new data point into the correct category in the future. This best decision boundary is called the hyperplane. SVM selects extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and thus the algorithm is called a support vector machine. Consider the diagram below in which there are two different categories that are classified using a decision boundary or hyperplane.



2.3 Logistic Regression

Logistic regression is one of the most popular machine learning algorithms, which falls under supervised learning techniques. It is used to predict a categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. So, the result must be a clear or distinct value. It can be either yes or no, 0 or 1, true or false, etc. But instead of giving an exact value as 0 and 1, it gives possible values that are between 0 and 1.

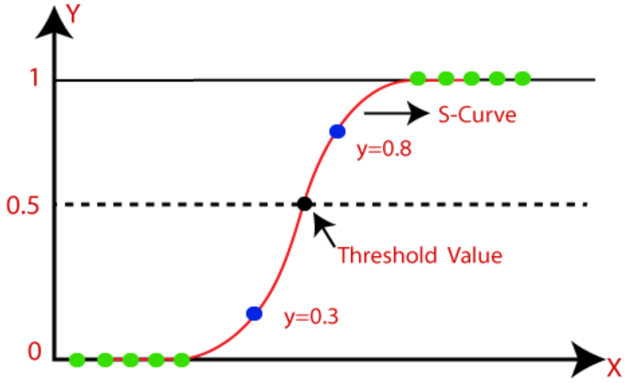
Logistic regression is similar to linear regression except for how it is used. Linear regression is used to solve regression problems, while logistic regression is used to solve classification problems.

In logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

A curve from a logistic function indicates the probability of something such as whether cells are cancerous or not, whether a mouse is obese or not based on its weight, etc.

Logistic regression is a significant machine learning algorithm because of its ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic regression can be used to classify observations using different types of data and the most effective variables used for classification can be easily determined. The following image shows the logistic function:



III. Literature Survey

The use of different ML algorithms allows early detection of many diseases such as heart, kidney, breast and brain diseases. Throughout the literature, the most commonly used prediction algorithms are SVM, RF, and LR, and accuracy is the most commonly used performance measure. The CN model has proven to be the most suitable for predicting common diseases. Furthermore, the SVM model showed better accuracy in kidney disease and PD due to its robustness when using high-dimensional, semi-structured and unstructured data. For breast cancer prediction, RF shows a greater advantage in the probability of correct disease classification due to its good scaling ability and tendency to avoid redundancy for large datasets. Finally, the LR algorithm proved to be the most reliable in predicting heart disease.

Future work should develop more sophisticated ML algorithms to improve disease prediction efficiency. In addition, the training model should be calibrated frequently after the training model to achieve better performance.

In addition, various demo graphics should be provided to the database to improve the accuracy and precision of the distributed model. Finally, more important feature selection techniques should be used to improve performance training models.

Marimuthu et al. focused on cardiovascular disease prediction using supervised ML method. The authors adjusted for data characteristics such as sex, age, chest pain, gender, purpose, and slope. The ML algorithms used are DT, KNN, LR and NB. According to the analysis, the LR algorithm gives the highest accuracy of 86.89% efficient compared to other mentioned algorithms. In 2018, Dwivedi tried to refine the prediction of heart disease by considering other parameters such as blood pressure, serum cholesterol in mg/dL and peak heart rate. The data collection used is from the UCI ML Lab; consists of 120 positive samples for heart disease and 150 negative samples for heart disease. Dwivedi tried to evaluate the performance of artificial neural network (ANN), SVM, KNN, NB, LR and classification tree. Using a 10-fold cross-sectional test, the results showed the highest classification accuracy and LR sensitivity.

It has high reliability in diagnosing heart disease. This result is confirmed by the results of Polaraj and Wahida et al. where logistic regression outperformed other methods such as ANN, SVM and Adaboost. This study managed to analyze the ML model comprehensively. For example, different hyperparameters are tested for each ML algorithm to achieve the best accuracy and precision. Despite this advantage, the small size of the obtained database limits the training model to target the disease with high precision and accuracy.

IV. Proposed Work

Limitations of this study are limited to the use of three supervised learning methods, Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT), to detect associations between heart disease data that can help improve prediction rates. The hardware and software requirements are as follows: We use Python, Machine Learning Tools, Spyder, Anaconda Navigator, StreamLit and Pickle as libraries. For diabetes, we propose a filtering method based on decision tree algorithm (Iterative Dichotomization 3) to select the most important features. Two ensemble learning algorithms, Ada Boost and Random Forest, are also used for feature selection, and we compare the performance of the classifier with an overlay-based feature selection algorithm. Several machine learning (ML) methods were performed with 2-fold, 5-fold and 10-fold recognition of voice features in Parkinson's disease. Principal component analysis (PCA) transformation function (FT) and k-nearest neighbor (k-NN) as classifiers with 10-fold cross-sectional test were found to have the best accuracy of 99.1%. A great machine learning web application that predicts disease. This presentation explains what preventable diseases and deaths are. He then reviews three separate documents to explain what has been done in the field and how the technology works. It will pave the way for future possibilities and enable disease prediction technology. A disease prediction system is a technology that can predict or diagnose a disease from a database and machine learning algorithm.

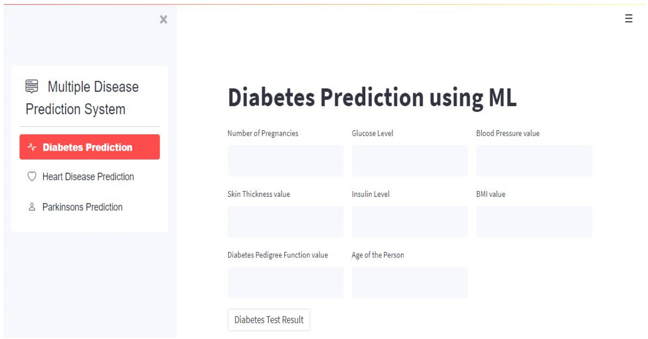


Fig.1. Web page of multiple Disease detection

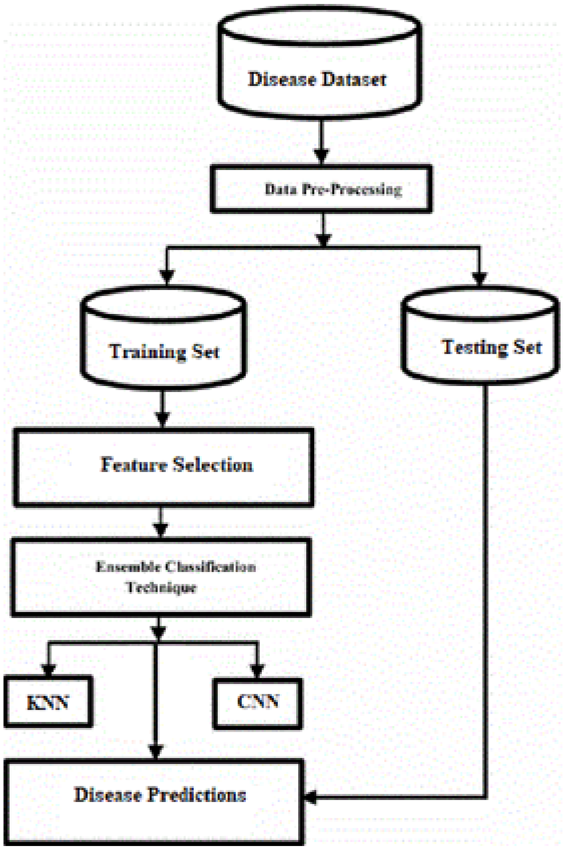


Fig.2. Block diagram of "Multiple Disease Detection Flowchart

First, the data for the project is collected and then divided into two parts. 80% training and 20% testing. First, 20,639 photos were taken. But after balancing, we got a total of 5000 images of different plant diseases like bacterial blight, leaf spots and healthy leaves. The deep learning model is then trained using a transfer learning method, and a training graph is extracted to show the meaning of the model. Then, performance metrics are used to classify the images, and finally, visualization techniques are used to detect and classify the images. The flowchart above illustrates this process. A basic CNN architecture is implemented. Applying the filter to the input image will result in a CNN feature map. Visualizing an element map means trying to understand a CNN.

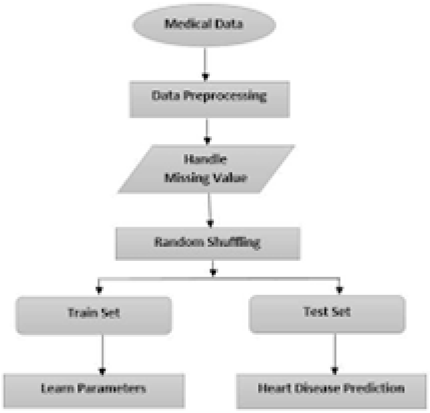


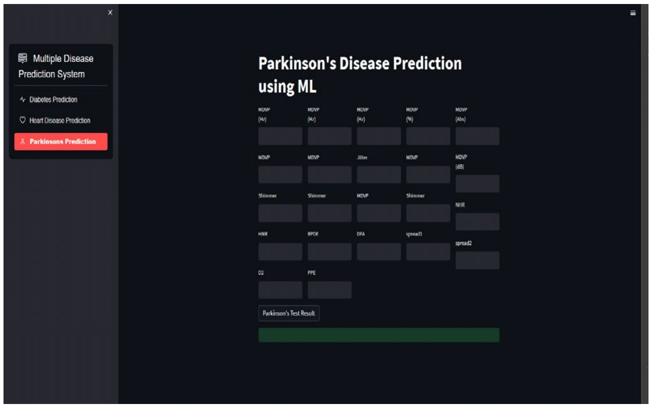
Fig.3. Flowchart of disease detection

4.1 Advantages

- It is very useful in hospitals
- Reduce workload of hospital staff.
- Easy to use. Anyone can use it.
- It predicts strong and accurate results.
- It takes less time to identify whether a patient has a particular disease or not.
- A needy patient can get early treatment by identifying the early stage of the disease.
- Cost savings.

V. Result Analysis

In systemic diabetes prediction model using knn algorithm, heart disease uses xgboost algorithm and liver uses random forest algorithm. So if the patient adds a parameter based on the disease, it indicates whether the patient has the disease or not. The parameter returns a range of required values and if the value is out of range, invalid or empty, a warning is displayed to add the appropriate value.



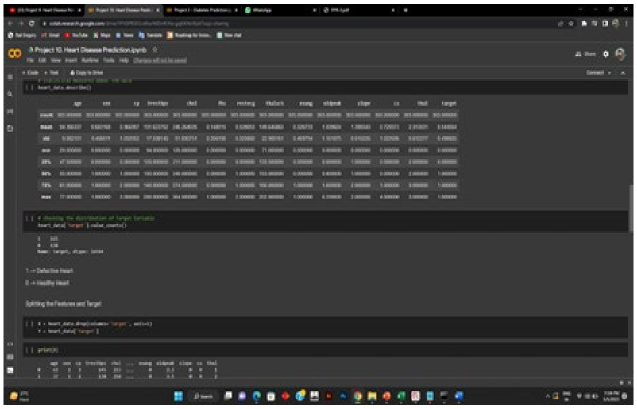
Parkinson's Disease Prediction using ML

MDVP (Hz)	MDVP (Hz)	MDVP (Hz)	MDVP (%)	MDVP (Abs)
MDVP	MDVP	Jitter	MDVP	MDVP (dB)
Shimmer	Shimmer	MDVP	Shimmer	NHR
HNR	RPDE	DFA	spread1	spread2
D2	PPE			

Parkinson's Test Result

Fig.4. Home Page of multiple disease detection system

- The proposed model is based on a disease detection system.
- The basic methodology of the system is to provide the patient with skin thickness, glucose level, BP level, insulin values, Age and Gender into the system's user interface, then it will predict diseases.
- Returns 1 if detected, 0 if not.


Project 10. Heart Disease Prediction.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
Copy to Drive
Connect
Share

```
heart_data.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
mean	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
std	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
min	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
50%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
75%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
max	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

```
# checking the distribution of Target Variable
heart_data['target'].value_counts()
```

```
1    165
0    138
Name: target, dtype: int64
```

1 --> Defective Heart

0 --> Healthy Heart

Splitting the Features and Target

```
X = heart_data.drop(columns='target', axis=1)
Y = heart_data['target']
```

```
print(X)
```

```
   age  sex  cp  trestbps  chol  ...  exang  oldpeak  slope  ca  thal
0   [illegible]
1   [illegible]
```

Fig.5. Output

Future Scope:

- We may add more diseases to the existing API in the future.
- We can try to improve the accuracy of prediction to reduce the mortality rate.
- Try to make the system user-friendly and provide a chatbot for general queries.
- Some studies use data from private hospitals. Efforts such as personal patient identification. information can be provided forget a big database. Given more data, advanced classifications will be more accurate. Because it's more information means more variety. As the model is trained on more examples, it becomes more general, reducing generalization errors.Medical information is hard to come by.

So, if the databases are made public, they can be accessed by researchers Additional Information.

VI. Conclusion

The main goal of this project is to develop a multi-disease prediction system with high accuracy. and This project saves users time from browsing different web pages. If the disease is detected early, it can be cured prolong your life and save you from financial problems. For this purpose, we use several machine learning algorithms Like Random Forest, XGBoost and Nearest Neighbor (KNN) to achieve the highest accuracy.

References

[1]. Agarwal, A., Chandrayan, S. and Sahu, S.S. (2016). “Prediction of Parkinson's disease using speech signal with Extreme Learning Machine,” in 2016 International Conference on Electrical, Electronic and Optimization Techniques (ICEEOT) (Chennai), 3776–3779.

[2]. Ali, L., Khan, S.U., Arshad, M., Ali, S. and Anwar, M. (2019a). "A multi-model framework for evaluating speech sample types with complementary information on Parkinson's disease," in 2019 International Conference on Electrical, Communication and Computer Engineering (ICECCE) (Swat), 1-5.

[3]. Anand, A., Haque, M.A., Alex, J.S.R. and Venkatesan, N. (2018). “Evaluation of Machine Learning and Deep Learning Algorithms Combined with Dimensionality Reduction Techniques for Parkinson's Disease Classification,” in 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (Louisville, KY), 342–347.

[4]. Animesh Hazra, Subrata Kumar Mandal, “Diagnosis and Prediction of Heart Disease Using Machine Learning and Data Mining Techniques: A Review”. Advances in Computational Sciences and Technology Volume 10, Number 7, 2017.

[5]. Ishaq Azhar Mohammed. (2019). A systematic literature mapping secure identity management using blockchain technology. International Journal of Innovations in Engineering Research and Technology, 6 (5), 86–91.

[6]. Hasan, A., Meziane, F., Aspin, R., & Jalab, H. (2016). “Segmentation of Brain Tumors in MRI Images Using Three-Dimensional Borderless Active Contouring. Symmetries,” 8(11), 132.

[7]. Anuradha S. Deshpande, Dhanesh D. Lokhande, Rahul P. Mundhe, Juilee M. Ghatol, "Lung cancer detection using CT and MRI image fusion using image processing". International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Number 3 March 2015.

[8]. Abbas Khosravi, Syed Moshfeq Salaken, , Amin Khatami, Saeid Nahavandi, Mohammad Anwar Hose "Lung Cancer Classification Using Deeply Acquired Features on a Low Population Dataset" IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE) 2017.