

Multiple Disease Detection Using Machine Learning

Mrs. Smiley Gandhi^{*1}, Prakhar Bhatt^{*2}, Harsh Srivastava^{*3}, Ayush Pandey^{*4}, Divyansh Barar^{*5}

^{*1}Assistant professor, Department of Computer Science & Engineering, BBDITM, Lucknow, UP, INDIA

^{*2,3,4,5}Student, Department of Computer Science & Engineering, BBDITM, Lucknow, UP, INDIA

ABSTRACT

Because of numerous contributing risk factors, including diabetes, high blood pressure, high cholesterol, irregular pulse rate, and many other factors, it is challenging to diagnose heart disease. The severity of cardiac disease in humans has been determined using various data mining and neural network techniques. According to the research paper the power of the proposed model was quite satisfactory and it was able to predict the evidence of heart disease in a particular individual using KNN and logistic regression which showed good accuracy compared to previously used classifiers like naïve bayes etc.

Healthcare is a very prominent research field with rapid technological advancement and increasing data day by day. In order to deal with a large volume of healthcare data we need Big Data Analytics which is an emerging approach in the Healthcare domain. Millions of patients seek treatments around the globe with various procedures. Making informed and effective decisions to enhance the general standard of healthcare will be aided by analyzing the trends in patient treatment for the diagnosis of a specific condition.

According to the research report, to forecast diabetes mellitus, we have used decision trees, random forests, and neural networks. As given in the paper Pima Indian Diabetes Dataset is also used for the prediction.

Diagnosis of Parkinson disease through a machine learning approach provides better understanding from PD dataset in the present decade. Orange v2.0b and weka v3.4.10 have been used in the present experimentation for the statistical analysis, classification, Evaluation and unsupervised learning methods. The Centre for Machine Learning and Intelligent Systems has retrieved a voice dataset for Parkinson's disease from the UCI Machine learning repository.

I. INTRODUCTION

Humans have always been susceptible to many illnesses and infections. Being healthy is essential for living a happy and wealthy life. We are developing a web application utilizing machine learning to assist users in recognizing many ailments, including Parkinson's disease, heart disease, and diabetes.

Parkinson's disease (PD) is a neurological ailment that affects many people. With an aging population, the therapy of Parkinson's disease is anticipated to become an increasingly essential and complex element of neurology and general medical practice.

Heart illness encompasses a wide range of issues affecting the heart. Because the heart is the most critical organ in the human body, heart disorders should not be taken lightly. Heart illnesses include coronary artery disorders, irregular heartbeats (arrhythmias), congenital heart abnormalities, heart muscle disease, and heart valve disease.

Diabetes mellitus is a common condition that affects how the body uses blood sugar (glucose). It is a common yet deadly illness. Type 1 diabetes, Type 2 diabetes, and gestational diabetes are the most common.

Using Machine Learning Techniques, the online application will assist a user in forecasting the existence or threat of the illnesses mentioned earlier. The user will need to submit specific data to anticipate the sickness. As we all know, we are entering the digital era, therefore this web application is an attempt to make things easier and more efficient for users and practitioners in making timely decisions on patients' health and treatment. For implementation, we shall use the Python programming language. The program will analyze a vast quantity of data to provide efficient and trustworthy findings.

II. ML TECHNIQUES FOR MULTIPLE DISEASE DETECTION

Supervised Learning Approaches:-

1. Support Vector Machine:-

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm that is commonly used for classification and regression tasks. It belongs to the family of discriminative models and has been widely adopted in various fields, including image recognition, text classification, and bioinformatics.

Here are the key components and concepts associated with SVM:

1. In SVM, the hyperplane is a decision boundary that separates the data points of different classes. For a binary classification problem, the hyperplane is a line in a two-dimensional space or a plane in a higher-dimensional space.
2. The margin is the distance between the hyperplane and the closest data points of each class. SVM aims to maximize this margin to achieve better generalization and robustness.
3. Support vectors are the data points that lie closest to the hyperplane and have the most influence on determining its position. They are crucial for defining the decision boundary and calculating the margin.

2. Neural Networks:-

Neural Networks, also known as Artificial Neural Networks (ANNs), are a class of machine learning models inspired by the structure and functioning of biological brains. Neural networks are highly flexible and powerful models that excel at pattern recognition, classification, regression, and other complex tasks.

At a high level, a neural network consists of interconnected nodes, called neurons or artificial neurons. These neurons are organized in layers, typically divided into three types:

1. **Input layer:** This layer receives the input data and passes it to the subsequent layers. The number of neurons in the input layer corresponds to the number of features in the input data.
2. **Hidden layer:** These layers are positioned between the input and output layers. Each neuron in a hidden layer receives inputs from the neurons in the previous layer, applies an activation function to the weighted sum of those inputs, and passes the result to the next layer.
3. **Output layer:** This layer produces the final output or predictions based on the information processed through the hidden layers.

3. Logistic Regression:-

Logistic Regression is a popular statistical and machine learning algorithm used for binary classification problems, where the goal is to predict the probability of an input belonging to a particular class. Despite its name, logistic regression is a classification algorithm and not a regression algorithm.

Some key characteristics and considerations of logistic regression include:

1. **Interpretability:** Logistic regression provides interpretable results as the coefficients associated with each feature indicate the direction and strength of the relationship between the features and the probability of the positive class.
2. **Linear decision boundary:** Logistic regression assumes a linear decision boundary, which means it may not capture complex nonlinear relationships between the features and the outcome without additional feature engineering or transformations.
3. **Regularization:** Regularization techniques such as L1 (Lasso) and L2 (Ridge) regularization can be applied to logistic regression to prevent overfitting and improve generalization by shrinking the coefficients.
4. **Assumptions:** Logistic regression assumes that the observations are independent, the relationship between the features and the log odds is linear, and there is no multicollinearity (high correlation) among the input features.

III. METHODOLOGY

Proposed Work:-

Diabetes

Today's globe is widely familiar with the term "diabetes," which poses serious problems for both industrialized and developing nations. Glucose from food can enter the bloodstream thanks to the hormone insulin, which is generated by the pancreas in the body. Diabetes is a lack of this hormone caused by pancreatic dysfunction, which can cause coma, kidney and retinal failure, pathological destruction of pancreatic beta cells, cardiovascular dysfunction, cerebrovascular dysfunction, peripheral vascular disease, dysfunction sex, joint failure, weight loss, ulcer, and pathogenic effects on immunity.

Heart Disease

Today, one of the main causes of death worldwide is heart disease. A significant difficulty in the study of clinical data analysis is the prediction of cardiovascular disease. In this article, we suggest a novel approach for identifying significant features using machine learning strategies that increase the precision of cardiovascular disease prediction. We achieve an enhanced performance level using a heart disease prediction model with an accuracy of 88.7% using a hybrid random forest and linear model.

Parkinson's Disease

Parkinson's disease is brought on by disturbance of the dopamine-producing brain cells, which are responsible for maintaining communication between brain cells.

The brain's dopamine-producing cells are what give movements control, adaptability, and fluidity. Parkinson's motor symptoms begin when 60 to 80% of these cells are destroyed because not enough dopamine is created.

IV. MODELING AND ANALYSIS

PROPOSED APPROACH:-

Heart Disease

The major goal is to estimate the proportion of people who have a good likelihood of having a heart (or cardiovascular) illness. To create this prediction system, we are utilizing Logistic Regression, a well-known machine-learning technique. Python is being used for implementation.

Logistic Regression

One of the most well-liked machine learning methods is supervised learning's logistic regression. When our data can be separated linearly and we need probabilistic outcomes, we typically employ this approach. Logistic regression is a statistical concept that is used in machine learning. It predicts a dependent variable based on one or more independent variables that were utilized to predict our desired outcome. It can be applied to multi-class classification as well as binary classification.

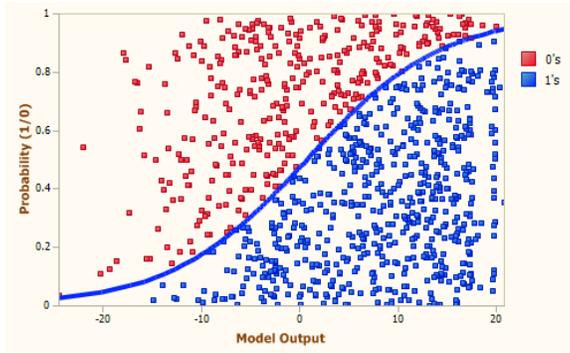


Fig 3.1: Model output for Logistic Regression Algorithm [8]

In Fig 3.1, it shows that the outcome or target variable in logistic regression is dichotomous. Dichotomous means there are only two possible classes. The workflow of the same algorithm is explained in Fig. 3.2.

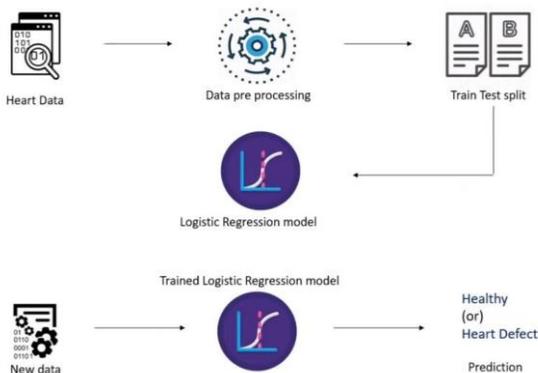


Fig 3.2: Workflow of Heart disease prediction system [9]

The basic outline of the whole process is laid in Fig 3.2, where we are able to see the steps involved in making this heart disease prediction system.

Diabetes disease

The major goal is to estimate the proportion of people who have a good risk of having diabetes. To create this prediction system, we are utilizing Support Vector Machine, a well-known machine learning technique. Python is being used for implementation.

Support Vector Machine

One of the most well-liked algorithms for supervised learning is called the Support Vector Machine (SVM), and it is used to solve both classification and regression issues. It is largely utilized in Machine Learning Classification issues, though. The SVM algorithm's objective is to establish the best decision boundary or line that can divide n-dimensional space into classes so that subsequent data points can be quickly assigned to the appropriate category. The term "hyperplane" refers to this optimal decision boundary. Fig. 3.3 has been used to describe the model results for the same.

Parkinson's Disease

The workflow comprises mostly Parkinson's data, which includes information on people with and without Parkinson's disease. The raw data is then pre-processed before being divided into training and testing data. This processed data is now

ready to be fed into our machine learning model, the Support Vector Machine Classifier, which will eventually tell us if a person has Parkinson's Disease or not. Fig. 3.5 has been used to demonstrate this.

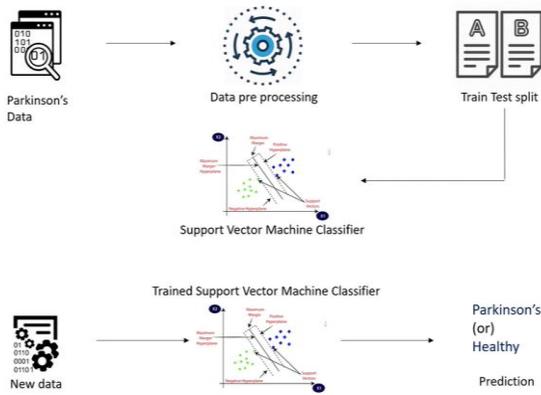


Fig 3.5: Workflow of Parkinson's disease prediction system [12]

Support Vector Machine

This is the model that will be utilized in the PD prediction system. Following data separation into training and testing data, the training data is drained into the model to train the model for the process. The testing data is used to evaluate the model once it has been trained. This is why the data has been divided into testing and training data. Fig. 3.6 has been used to demonstrate this. Essentially, the model determines the hyperplane that distinguishes those who are suffering from those who are not by analyzing data based on distinct patterns.

V. RESULTS AND DISCUSSION

All of the investigations in this literature for Diabetes Type II were carried out under identical experimental settings using the Pima Indian Diabetes Dataset. The models were examined using five-fold cross-validation in this study. To validate the approaches' general applicability, we picked several methods with superior performance for conducting independent test studies. In medicine, diabetes is diagnosed based on fasting blood glucose, glucose tolerance, and random blood glucose readings. Based on daily physical exam data, machine learning can assist people in developing a preconception about diabetes and serve as a reference for clinicians.

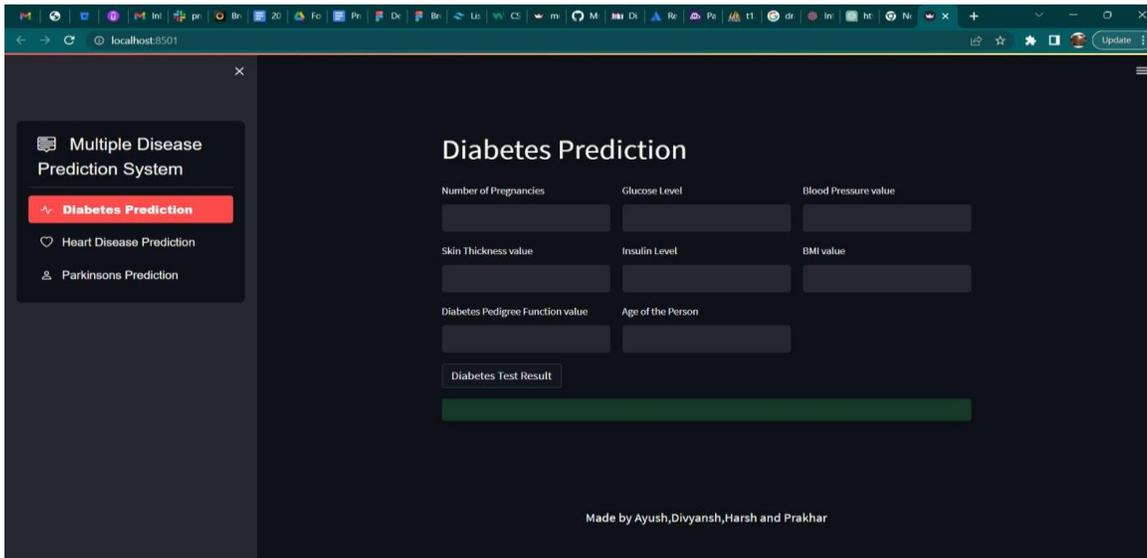
To forecast cardiac disease, many strategies have been developed to abstract away knowledge using existing data mining methods. This approach successfully identifies the patient's condition by using several clinical characteristics such as left bundle branch block, right bundle branch block, atrial fibrillation, normal sinus rhythm, sinus bradycardia, atrial flutter, premature ventricular contraction, and 2nd-degree block. connected to heart disease. The suggested model's predictive ability was pretty excellent, and it was able to predict the presence of heart disease in a specific individual using KNN and logistic regression.

In the case of Parkinson's disease (PD), The recovered dataset demonstrates statistical analysis, classification, evaluation, and the provision of unsupervised learning methods. Methods for determining the characteristics that are likely to indicate

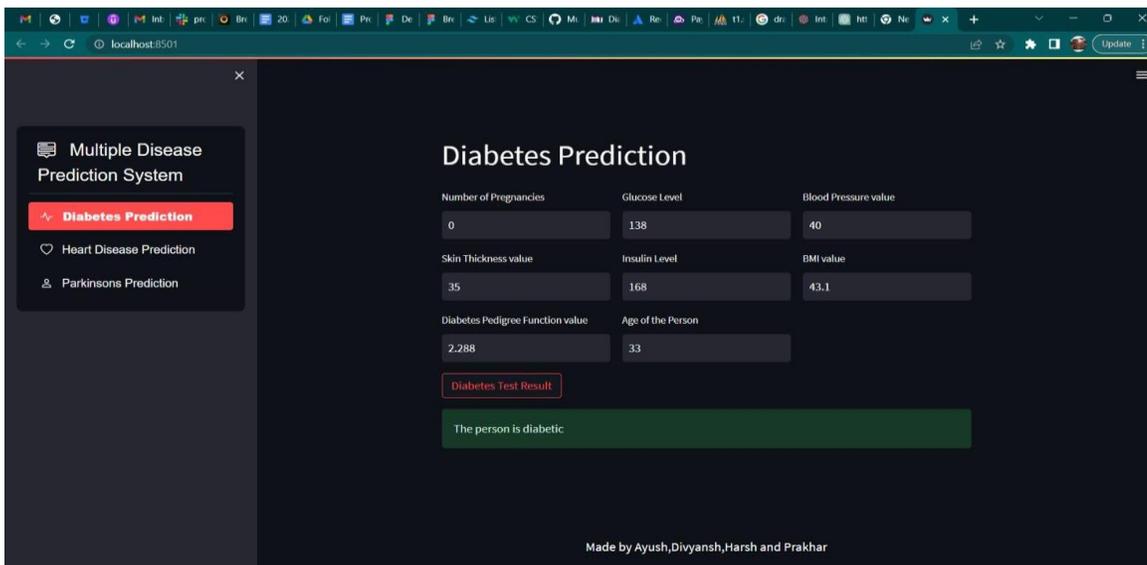
the existence of Parkinson's disease. In the current decade, speech data analysis is critical for understanding and developing diagnostic tools for human disorders. The current technique diagnoses Parkinson's disease using a speech dataset and machine learning algorithms.

VI. WORKING MODEL

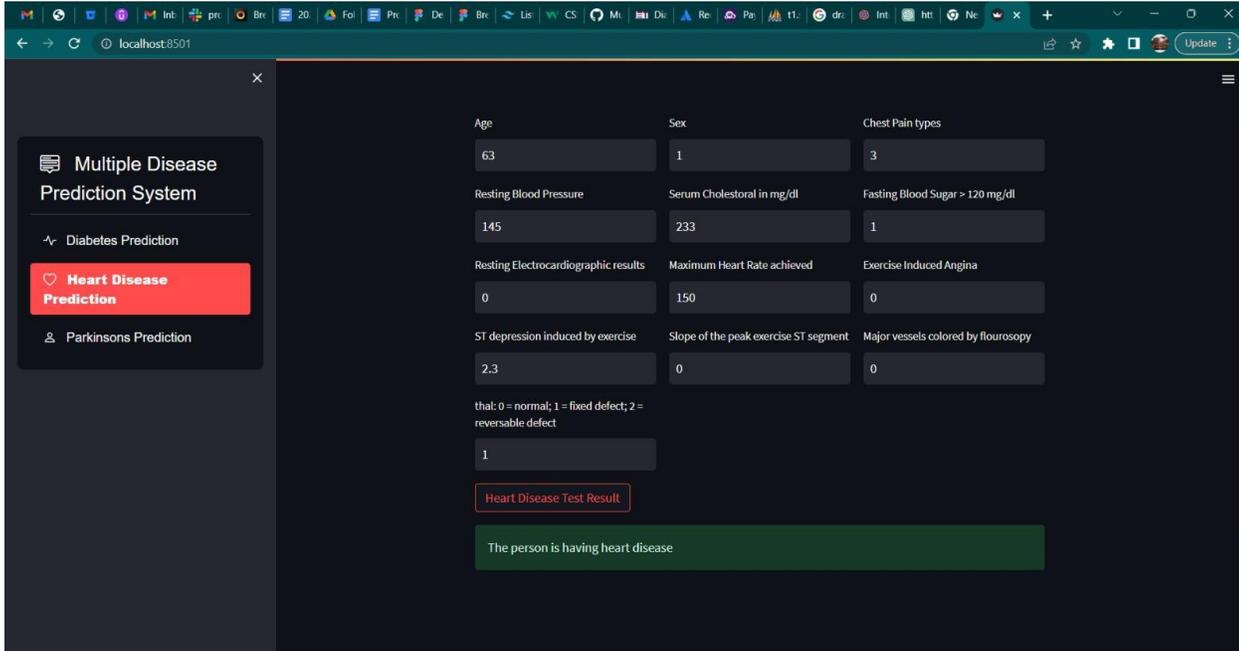
1. User Interface:-



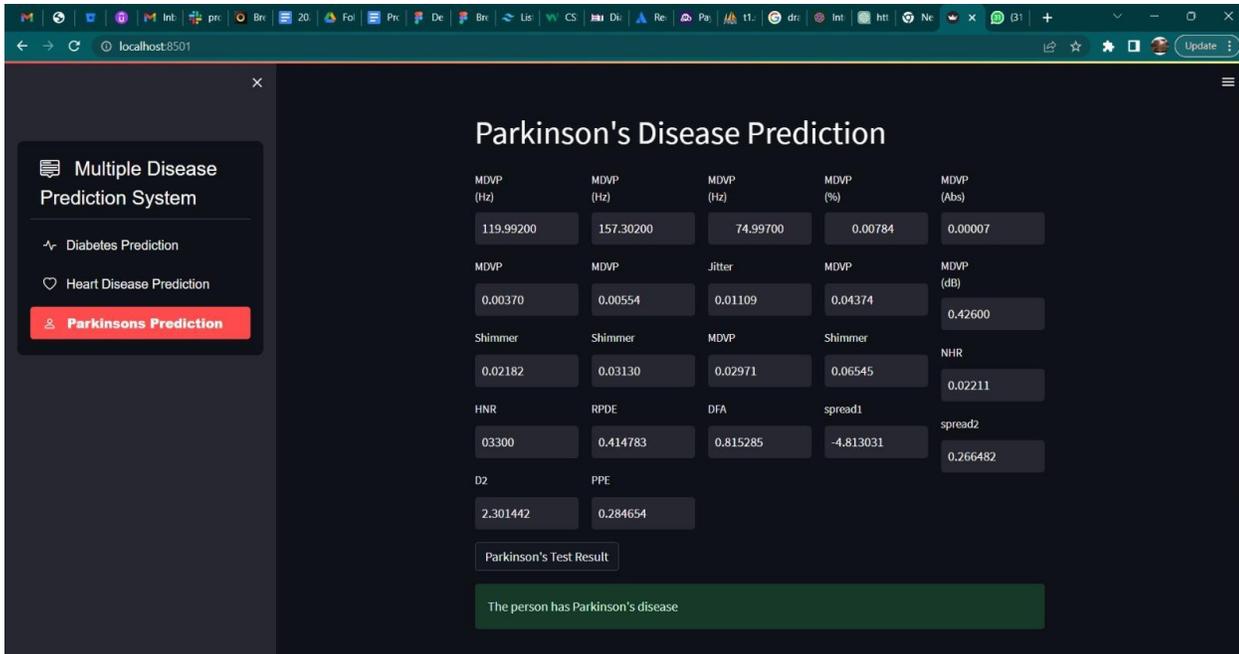
2. Person suffering from Diabetes:-



3. Person suffering from Heart Disease:-



4. Person suffering from Parkinson's Disease:-



VII. FUTURE WORK

In the future, by employing several machine learning approaches to forecast the existence of illnesses, we will be able to provide a prediction performance report for each methodology, allowing us to optimize our prediction algorithm.

As we all know, we generate a large amount of data every day; the more data that is given into the database, the more intelligent the system becomes, and therefore our accuracy improves.

To expand our services, we want to include many more sections including more ailments that can be predicted using machine learning algorithms. Our objective is to create a platform where people can come and anticipate many diseases, making it a one-stop shop for the target population.

The present COVID-19 epidemic has made the entire globe sit up and take notice because medical catastrophes are unexpected and can bring severe financial disruption. That is a big reason why, in today's contemporary world, people are highly health concerned and the majority of them look forward to purchasing some type of medical insurance, thus we are looking forward to integrating a medical cost prediction system into our program to tackle that problem.

Finally, because we all know that "prevention is better than cure," we might include a part in our application emphasizing the value of good behaviors in preventing illnesses and living a healthy life.

VIII. CHALLENGES

1. Data Quality and Availability:-

In the context of multiple disease detection using machine learning, data quality and availability are critical factors that can significantly impact the performance and generalizability of the models. This section explores the challenges associated with data quality and availability and discusses potential strategies to address these issues.

2. Interpretability and Explainability:-

Interpretability and explainability are crucial aspects of machine learning models in the context of multiple disease detection, particularly in healthcare. Healthcare professionals and patients require transparent and understandable models to trust and effectively utilize the predictions made by these models. This section explores the challenges associated with interpretability and explainability and discusses potential strategies to enhance these aspects.

3. Ethical considerations and Privacy concerns:-

The use of machine learning for multiple disease detection in healthcare raises important ethical considerations and privacy concerns. It is crucial to address these issues to ensure responsible and ethical implementation of machine learning models.

IX. CONCLUSION

In this prediction system, we created a platform for individuals to anticipate their specific condition and discover a remedy based on it.

We used two important machine learning techniques or algorithms, namely logistic regression and support vector machine, to create a more precise prediction system by analyzing different patterns in the processed data that have been drained into the machine learning model for respective diseases.

Based on these two methods, we have a graph with a positive and a negative hyperplane, indicating whether or not a person is suffering from the condition. Aside from that, we've established an awareness and prevention area where people can learn more about the condition and how to avoid it.

In terms of design, we created an engaging user interface for users that would pleasure them and make things simpler for individuals of all ages.

X. REFERENCES:-

- [1] Zou Q, Qu K, Luo Y, Yin D, Ju Y, and Tang H (2018) Predicting Diabetes Mellitus With Machine Learning Techniques. *Front. Genet.* 9:515. doi: 10.3389/fgene.2018.00515
- [2] Md Kamrul Hasan, Md Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan, "Diabetes prediction using ensembling of different machine learning classifiers", *IEEE Access*, vol. 8, pp. 76516-76531, 2020.
- [3] Harshit Jindal et al 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* 1022 012072
- [4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", *IEEE Access*, vol. 7, pp. 81542-81554, 2019.
- [5] T. V. Sriram, M. V. Rao, G. S. Narayana, D. S. Kaladhar, and T. P. Vital, "Intelligent Parkinson disease prediction using machine learning algorithms", *IJEIT.*, vol. 3, pp. 212–215, 2013.
- [6] Z. Karapinar Senturk, *Early Diagnosis of Parkinson's Disease Using Machine Learning Algorithms, Medical Hypotheses*, 2020.
- [7] Timothy J. Wroge, Yasin Özkanca, C. Demiroğlu, Dong Si, David C. Atkins and R. Ghomi (2018), "Parkinson's Disease Diagnosis Using Machine Learning and Voice", *Computer Science, IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*.
- [8] "Customer Subscription Analysis And Prediction Based On App Behavior Analysis Logistic Regression", Shekhar Koirala, *Medium*, 2019.
- [9] Fig. 2, Kavya Sreehari, Devika Santhosh Kumar, Muhammed Shameem S, Sumeesh S, and Bismi M, "Prediction of Heart Disease Using Logistic Regression", 2022, *International Research Journal of Engineering and Technology (IRJET)*.
- [10] "Support Vector Regression and its Mathematical Implementation", Rahul Rastogi, *Medium*, 2020.
- [11] "Diabetes Prediction using Machine Learning with Python", 2021, Siddhardhan, *YouTube*.

- [12] “Parkinson’s Disease Detection using Machine Learning - Python”, 2021, Siddhardhan, YouTube.
- [13] “Ramani, V. R., Prasad, B. V. V., & Das, M. (2019). Parkinson's Disease Diagnosis using Machine Learning Algorithms”
- [14] ”Basak, J. (2021). Detection and Prediction of Parkinson's Disease using Machine Learning.”
- [15] “Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks.”
- [16] “Cho, Y., Park, H., Choi, Y., & Kim, J. (2019). Cardiac arrhythmia detection from 12-lead electrocardiogram using attention-based convolutional and long short-term memory networks.”
- [17] “Jain, V., & Raina, S. (2020). Diabetes Mellitus Prediction Model Using Machine Learning Techniques. Journal of Healthcare Engineering, 2020”
- [18] “Alsmadi, M. K., & Malaekah, H. M. (2020). Diabetes Disease Prediction Using Machine Learning Techniques.”
- [19] “Tiwari, S., Sharma, V., & Dhawale, P. (2019). Prediction of Diabetes Mellitus Using Machine Learning Techniques.”