Advanced Health Prognosis Tool

Ms.V.P.V.Bharathi[1], Mr. Sai Ram Terli[2], Ms. Vennela Lalam[2], Mr. Mohammed Anish[2] and Mr. Ch Adit Rohan[2]

[1]*Assistant professor, Dept. of CSE, Raghu Engineering College, Dakamarri(V), Bheemunipatnam, Visakhapatnam District, 531162*

[2]*Department of Data Science, Raghu Engineering College, Dakamarri(V), Bheemunipatnam, Visakhapatnam District, 531162*

***Abstract*— Chronic diseases such as Heart disease, Diabetes, and Parkinson's disease etc. are leading causes of mortality in India and globally, underscoring the urgent need for accurate diagnostic tools to address these conditions. This study presents a machine learning (ML)-based diagnostic system aimed at enhancing disease prediction accuracy and enabling early intervention. The proposed model analyzes a dataset comprising patient records diagnosed with the diseases, utilizing relevant symptoms to optimize predictions. Employing various machine learning algorithms—including SVM, classification algorithms—our system achieves a high prediction accuracy. The integration of extensive medical data and advanced techniques enables effective analysis for early disease identification, patient care, and community health services. Additionally, the model's ability to handle both structured and unstructured data enhances its predictive capabilities for the diseases. By leveraging Stream lit, the system facilitates real-time disease prediction, providing healthcare practitioners with valuable insights to improve patient outcomes and reduce mortality rates. A key feature is the integration of an AI-driven chatbot powered by OpenAI, which provides users with an interactive conversational interface. The chatbot assists users by answering health-related inquiries, making it a valuable tool for individuals seeking timely health information and support.**

***Keywords*— Prediction, Randomforest, Decision Tree, SVM Classifier, Exploratory Data Analysis, Machine Learning, Deep Learning.**

1. INTRODUCTION

The Project "Advanced Health Prognosis Tool" is a revolutionary stride in the realm of predictive healthcare. This innovative system represents the harmonious blend of cutting-edge technology and advanced machine learning algorithms, all working in unison to transform the landscape of disease prediction and diagnosis. Here's a deeper dive into its key aspects:

Core Objectives and Impact:

* The tool is designed with the primary aim of accurately predicting diseases at an early stage. It's a proactive approach to healthcare that anticipates health issues before they escalate.
* By identifying potential health risks early, the system plays a crucial role in enhancing patient outcomes. Early detection often leads to more effective treatment plans.
* It's also about making healthcare management more efficient. By providing timely and accurate predictions, it helps in reducing the overall burden on healthcare systems.

Technological and Methodological Sophistication:

* At its heart, the system uses sophisticated machine learning algorithms. These algorithms are adept at analyzing vast datasets, identifying complex patterns, and predicting diseases with remarkable accuracy.
* The tool's versatility is one of its key strengths. It's capable of predicting a wide array of diseases, showcasing its adaptability and comprehensive nature.
* The integration of data-driven insights and machine learning marks a significant paradigm shift in healthcare. This approach not only redefines disease prediction but also significantly improves patient care.

User-Centric Design and Accessibility:

* The system is designed to be incredibly user-friendly. It features an intuitive interface that makes it easy for users to navigate and understand.
* Accessibility is a key focus. The tool is designed to be available to a wide range of users, ensuring that advanced healthcare insights are not confined to a select few.
* By offering proactive insights of health, the system empowers individuals to take control of well-being. It's a tool that enables informed and healthcare decisions.

Development and Technology Integration:

* The development of this tool is grounded in the principles of data-driven healthcare and predictive analytics, ensuring a robust and reliable system.
* It leverages Python-based tools and libraries, including Scikit-learn and Streamlit, showcasing the use of cutting-edge technology in healthcare.
* The aim is to create a platform that not only predicts diseases but also offers personalized healthcare recommendations, tailored to individual needs and preferences.

The *"Advanced Health Prognosis Tool"* is more than just a technological innovation. It's a holistic approach to healthcare that emphasizes early detection, personalized care, and proactive health management. By doing so, it not only improves patient outcomes but also contributes to building a healthier and more informed society.

2. LITERATURE REVIEW

1.According to the paper, diabetes is one of the dangerous diseases in the world, it can cause many varieties of disorders which includes blindness etc. In this paper they have used machine learning techniques to find out diabetes disease as it is easy and flexible to forecast whether the patient has illness or not . Their aim of this analysis was to invent a system that can help the patient to detect the diabetes disease of the patient with accurate results. Here they used mainly 4 main algorithms Decision Tree, Naïve Bayes , and SVM algorithms and compared their accuracy which is 85%,77%, 77.3% respectively . They also used ANN algorithm after the training process to see the reactions of the network which states whether the disease is classified properly or not. Here they compared the precision recall and F1 score support and accuracy of all the models.[1]

2. The main aim of the paper is, as the heart plays an important role in living organisms. So, the diagnosis and prediction of heart related disease should be perfect and correct because it is very crucial which can cause death cases related to heart .So, Machine learning and Artificial Intelligence supports in predicting any kind of natural events .So in this paper they calculate accuracy of machine learning for predicting heart disease using k-nearest neighbor ,decision tree, linear regression and SVM by using UCI repository dataset for training and testing . They also compared the algorithm and their accuracy SVM 83 %,Decision tree 79%,Linear regression 78%,k-nearest neighbor 87%[2]

3. The system defines that liver diseases are causing a high number of deaths in India and is also considered as a life threatening disease in the world. As it is difficult to detect liver disease at an early stage. So using automated programs using machine learning algorithms we can detect liver disease accurately .They used and compared SVM ,Decision Tree and Random Forest algorithms and measured precision, accuracy and recall metrics for quantitative measurement. The accuracy is 95%,87%,92% respectively.[3]

A lot of analysis over existing systems in the health care industry considered only one disease at a time. For example, one system is used to analyze diabetes, another is used to analyze diabetes retinopathy, and another system is used to predict heart disease. Maximum systems focus on a particular disease. When an organization wants to analyze their patient's health reports then they have to deploy many models. The approach in the existing system is useful to analyze only particular diseases. Some of the models have lower accuracy which can seriously affect patients' health. When an organization wants to analyze their patient's health reports, they have to deploy many models which in turn increases the cost as well as time Some of the existing systems consider very few parameters which can yield false results.

3. MATERIALS AND METHODS

3.1 Data Pre-processing:

For a system to predict proper results, first it should be trained properly with existing data. Pre-Processing the data is important so that good quality data is used for training the model. Data cleaning and removal of noise are some of the processes involved in Pre Processing. We used the Diabetes Dataset of UCI Machine learning Repository, For heart disease analysis Cleveland, heart disease patient's data sets are used. And for Parkinsons Data Set which is available in machine learning repository. Data from various sources has been collected and aggregated. Now by using the preprocessing techniques like:

Data Cleaning:
Data is cleansed through processes such as filling in missing value, thus resolving the inconsistencies in the data.
Data Reduction:
The analysis becomes hard when dealing with a huge

database. Hence, we eliminate those independent variables(symptoms) which might have less or no impact on the target variable(disease). In the present work, 95 of 132 symptoms closely related to the diseases are selected.

3.2 Model Selected

The system is trained to predict the diseases using three algorithms Decision Tree Classifier, Randomforest Classifier & SVM Classifier A comparative study is presented at the end of work, thus analyzing the performance of each algorithm of the considered database. Algorithm with the best accuracy and precision will be considered.

3.3 Random Forest Classifier

Random forest is an adaptable, simple to use machine learning algorithm that provides exceptional outcomes more often than not even without hyper tuning. The major limitation of decision tree algorithms is overfitting. It appears as if the tree has memorized the data. Random Forest prevents this problem: It is a version of ensemble learning. Ensemble learning refers to using multiple algorithms or the same algorithm multiple times.

Random forest is a group of Decision trees. And greater the number of these decision trees in Random forest, the better the generalization. More precisely, Random forest works as follows:
1. Selects k symptoms from dataset (medical record) with a total of m symptoms randomly (where k<<m). Then, it builds a decision tree from those k symptoms.
2. Repeats n times so that we have n decision trees built from different random combinations of k symptoms (or a different random sample of the data, called bootstrap sample).
3. Takes each of the n-built decision trees and passes a random variable to predict the Disease. Stores the predicted Disease, so that we have a total of n Diseases predicted from n Decision trees.
4. Calculates the votes for each predicted Disease and takes the mode (most frequent Disease predicted) as the final prediction from the random forest algorithm.

3.4 SVM Classifier

The full form of SVM is Support Vector Machine. SVM is a supervised machine learning algorithm capable of performing classification, regression, and even outlier detection. It is mostly used for classification problems. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N-the number of features) that distinctly classifies the data points.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, SVM chooses the extreme vectors that help create the hyperplane. These extreme cases are called support vectors, and hence the Support Vector Machine.

Let's consider a simple scenario: classifying data into two categories. Visually, you can think of this as plotting data points on a graph and drawing a line between them to divide them into two groups. In SVM, this line is our hyperplane.
Finding the Right Hyperplane (Scenario 1)

* In Scenario 1, both Hyperplane A and Hyperplane B can perfectly classify the data. However, which one is the optimal choice?
* The answer lies in selecting the hyperplane farthest from the nearest data points of both classes. These nearest data points are known as *support vectors*.
* The distance between the hyperplane and the support vectors is called the margin. The best hyperplane or optimal hyperplane is the one with the maximum margin.

Finding the Right Hyperplane (Scenario 2)

* In Scenario 2, data points are scattered in such a way that it's impossible to draw a straight line (hyperplane) to separate them.
* SVM tackles this by introducing a *kernel trick*. It maps the data into a higher-dimensional space where finding a hyperplane becomes possible.

SVM Uses the Kernel Trick

* The kernel trick takes data from the original space and transforms it into a higher-dimensional space.
* In this new space, it's possible to find a hyperplane that beautifully separates the classes.

SVM: Choosing the Right Parameters

* The effectiveness of SVM depends on selecting the right kernel and parameters.

* It's crucial to tune these parameters carefully to achieve the best performance.

3.5 Model Building

Multiple disease prediction is a classification problem. So, we have implemented various classification algorithms like Decision Tree Classifier, Random Forest Classifier, SVM, to choose the algorithm with best results. Next step is building the prediction version. First we need training and testing of the data by using the classification algorithm then we are fitting the model and load the model by using a pickle module.. Second, by using the python module Streamlit we are developing the user interface where the user is able to see the prediction of disease. Algorithm is chosen with the present stage of version accuracy.

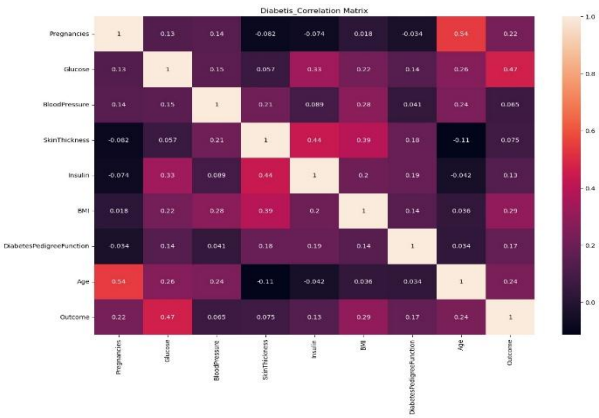
4. RESULTS AND DISCUSSION

4.1 The confusion matrix provides a performance evaluation of a classification model by comparing predicted labels with actual labels. It consists of four key values: True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP) The model correctly classified instances as "No" (negative class) and instances as "Yes" (positive class), demonstrating strong accuracy. However, cases were incorrectly classified as "Yes" when they were actually "No" (FP), and case was wrongly classified as "No" when it was actually "Yes" (FN). The low number of misclassifications suggests that the model performs well in distinguishing between the two classes, with a high sensitivity (recall) for detecting the positive class and relatively few false predictions.

4.2 The AI-powered chatbot in this disease prediction system acts as a virtual health assistant, bridging the gap between users and intelligent healthcare support.

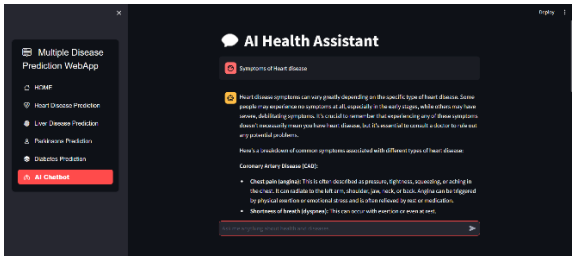
It enables real-time interaction, allowing users to describe their symptoms and receive preliminary insights based on machine learning-driven analysis. By utilizing natural language processing (NLP), the chatbot interprets user inputs and provides relevant health information, lifestyle recommendations, and potential risk assessments. Additionally, it enhances accessibility by guiding users on whether to seek medical attention, making healthcare more proactive. Its seamless integration with predictive models ensures that users receive AI-assisted consultation, improving overall engagement and decision-making in early disease detection.

4.3 The correlation heatmap visualization serves as a powerful tool for exploring relationships between multiple variables within our dataset. Using Seaborn's heatmap functionality (sns.heatmap), we created a color-coded matrix displaying correlation coefficients between all numeric features, providing an intuitive visual representation of data patterns. The vibrant color spectrum—ranging from deep blue (strong negative correlations) to intense red (strong positive correlations)—allowed us to quickly identify significant feature relationships that might otherwise remain hidden in tabular data. This technique proved particularly valuable for feature selection processes, highlighting potential multicollinearity issues between variables that could affect model performance. By adjusting parameters such as color palettes, annotation formats, and masking options, we customized the visualization to emphasize the most statistically significant relationships. The resulting heatmap not only guided our feature engineering decisions but also facilitated clear communication of complex data interdependencies to non-technical stakeholders. This visualization technique ultimately helped optimize our model by ensuring we focused on truly informative variables while avoiding redundancy in our feature space.



Classification Report:

Accuracy Scores	Accuracy on Heart Disease	Accuracy on Diabetes	Accuracy on Parkinsons	Accuracy on Liver	Accuracy on Pneumonia
SVM	0.98	-	0.96	-	-
Decision Tree	-	-		0.93	-
Random Forest	-	0.99	-	-	-
CNN	-	-	-	-	0.96



5. CONCLUSION

The "Advanced Health Prognosis Tool" represents a significant stride in leveraging technology to enhance healthcare outcomes. This tool, which uses machine learning algorithms on extensive datasets, can predict the likelihood of various diseases, including heart disease, diabetes, and cancer, with remarkable accuracy. The tool's development underscores the increasing role of machine learning in predictive medicine, demonstrating its potential to revolutionize how we approach disease diagnosis and management.

A critical aspect of this tool is its emphasis on early detection and intervention. By analyzing a wide array of patient data, it can identify patterns and risk factors that may not be immediately apparent through traditional diagnostic methods. This capability is invaluable in the context of diseases like cancer, where early detection can significantly improve treatment outcomes. Moreover, the tool's ability to predict the likelihood of chronic conditions such as diabetes and heart disease can empower individuals to make timely lifestyle adjustments, potentially preventing the onset or progression of these diseases.

Accessibility and user-friendliness are central to the design of the Advanced Health Prognosis Tool. Its intuitive interface ensures that individuals can easily input their data and receive clear, understandable results. This design philosophy democratizes access to advanced healthcare insights, making them available not just to medical professionals but also directly to individuals. Furthermore, the tool's integration with electronic health records streamlines the data input process, enhancing its practicality in real-world healthcare settings.

However, the tool also brings forth considerations, particularly concerning data privacy and the potential for over-reliance on technology. As it processes sensitive personal data, ensuring stringent data protection and Compliance with ethical standards are paramount. Additionally, while the tool offers valuable predictive insights, it is essential to remember that it complements, not replaces, human medical expertise. The nuanced understanding and clinical judgment of healthcare professionals remain indispensable in patient care.

In conclusion, the Advanced Health Prognosis Tool is a significant advancement in Medical Technology. Its ability to accurately predict a range of diseases, coupled with its focus on early detection and user-friendly design, marks a significant step forward in proactive and personalized healthcare. As technology continues to evolve, tools like these will likely become increasingly integrated into healthcare systems, offering the potential to improve patient outcomes on a global scale.

6. REFERENCES

[1] M. Kalpana Chowdary , Rajsekhar Turaka , K. Anil Kumar , B. Devananda Roa C. Ganesh (2023). " "Multiple Disease Prediction by Applying Machine Learning and Deep Learning Algorithms.".

[2] Laxmi Deepthi Gopisetti ,Srinivas Karthik Lambavai Kummera ,Sai Rohan Pattamsetti, Sneha Kuna, Niharika Parsi , Hari Priya Kodali (2023). "Multiple Disease Prediction System using Machine Learning and Streamlit".

[3] Mahendran K., Surya S., Thejashrayal E (2023). "Streamlit-Powered Comprehensive Health Analysis and Disease Prediction System".

[4] Sayali Ambekar, Rashmi Phalnikar (2022). "Disease Risk Prediction by Using Convolutional Neural Network".

[5] Sneha Grampurohit, Chetan Sagarnal (2022). "Disease Prediction Using Machine Learning Algorithms".