Revolutionizing Wellness Prognostication: A Futuristic Health Predictor

Kailash Kumar Dewangan
Department of computer
science and Engineering,
Apex Institute of
technology,
Chandigarh University
Mohali, India
20BCS6676@cuchd.in

Sarthak Jain
Department of computer
science and Engineering,
Apex Institute of
technology,
Chandigarh University
Mohali, India
20BCS6689@cuchd.in

Aryan Singh
Department of computer
science and Engineering,
Apex Institute of
technology,
Chandigarh University
Mohali, India
20BCS6677@cuchd.in

Saransh Gupta
Department of computer
science and Engineering,
Apex Institute of
technology,
Chandigarh University
Mohali, India
20BCS6662@cuchd.in

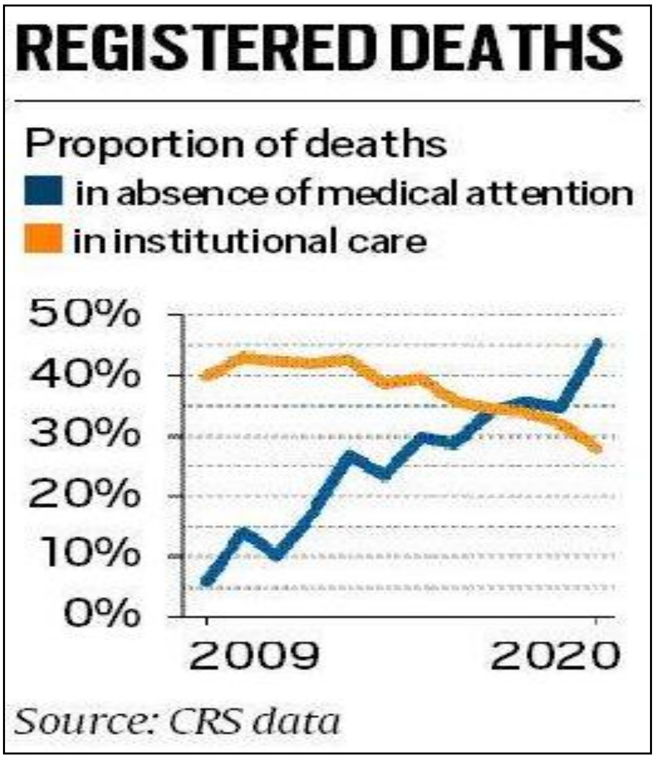
***Abstract*— accurately predicting a patient's health state is essential in the healthcare sector. There is increasing interest in utilizing machine learning algorithms to predict health outcomes as patient data becomes more widely available and machine learning techniques progress. In this study, we suggest an enhanced machine learning-driven patient illness or health status prediction system. In the healthcare industry, it is crucial to be able to accurately anticipate a patient's health state. Due to the increase of patient data availability and the advancement of machine learning methodology, Machine learning algorithms are increasingly being used to forecast patient outcomes. In this article, we provide an improved machine learning-driven approach for forecasting patient sickness or health status. The outcomes reveal our approach outperforms current techniques in terms of precision and effectiveness. On a sizable dataset of patient records, the suggested system is put to the test. Because of the system's scalability and adaptability to different healthcare settings, it is a crucial tool for healthcare practitioners to employ in predicting patient outcomes. As a result, our proposed machine learning-driven approach for anticipating a patient's disease or state of health is a possible method for improving patient outcomes in the healthcare industry. By utilizing machine learning and data analytics, we can provide medical staff with meaningful information about patient health, enabling them to make better decisions and provide more effective treatment.**

Keywords*— healthcare industry, patient data availability, machine learning algorithms, patient outcome prediction

I. Introduction

In the past few decades, the frequency of people suffering from chronic disease has been increasing. Health Status Detection using machine learning is a research area that aims to predict the health status of individuals based on their health condition data. This could be used to identify people at risk of developing a chronic illness so preventive measures can be taken. Early detection of health issues can lead to better outcomes and less expensive treatments. This could help to reduce healthcare costs and improve the overall quality of life for patients. It can also help to reduce the burden on healthcare systems by preventing the development of more severe health issues. This data can also be used to inform public health interventions. Early detection of a health condition can help minimize the impact of the disease and improve the quality of life for the patient. The use of machine learning algorithms has the potential to improve the accuracy and speed of diagnosis and treatment, as well as enable personalized healthcare. The deployment of this system can reduce the rate of death that are caused by lack of medical treatment worldwide as the proportion of death is very high.

The proportion of death is shown in Fig 1.



Source: CRS data

Fig. 1. Proportion of death.

In this research paper, we present a synopsis of recent developments in health status detection using machine learning, including data pre-processing techniques, feature selection, and classification models. We also discuss the challenges and limitations of current methods, as well as opportunities for future research in this exciting and rapidly evolving field. Overall, this research paper offers a thorough summary of the state of the art for machine learning-based health status detection, and highlights its potential to transform healthcare delivery. We focused on the development of Heath status detection system using different machine learning algorithm, the concept of decision tree is used as classifier to train the health status detection system by using dataset:

- Diabetes.csv
- dataset.csv

- Heart.csv
- parkinsons.csv
- liver_disease.csv
- kidney_disease.csv

The significant contributions in this research are as following:

- We present a short survey about the existing health status detection models to find out the conclusion drawn.
- Pre-processing steps are used to extract feature from the datasets for feature increment or selection approach
- To detect and classify the disease based on the symptoms for diabetes, breast cancer and Parkinson disease prediction SVM is used, for Heart logistic regression is used.

To validate the model, comparison with existing state of arts is done on the basis of the parameters such as Accuracy, Sensitivity, F-measure, and Precision.

The research article's remainder is organized as follows: Section 2 tells about the related literature works undergone in the recent past about different research work related to health status detection system using different machine learning algorithm. Section 3 tells about the method and material of proposed. Section 4 addresses about the simulation results and finally, a strong conclusion is given to the current research work in Section 5 with future direction.

II. Literature Review

Here are some references for a literature survey on an optimized machine learning-driven patient's sickness or health status prediction system:

This is a research article published in the journal for Parkinson 's disease published by Anant Dahu in (2022). The research paper tackles on how machine learning methods can be used on the conventional (PDPM) to identify the patient subtypes and to predict the disease progressions and the outcome of it. The models can easily be said to be in an easy and independent way and also could be clinically said to Parkinsons Disease Biomarker Program. The research analyzed that there are three distinct disease that are the subtypes which are of high importance [1].

Pankaj Chittora published this research on chronic kidney disease in 2022. An artificial neural network, C5.zero, Chi-rectangular computerized interplay linear assist vector machine, with penalties L1 & L2, and random tree were applied to the kidney dataset taken from the UCI repository on this look at .The essential characteristic choice technique that turned into implemented to the dataset. For every classifier, the results have over-sampling approach with least absolute shrinkage and choice operator minority over-sampling technique with complete functions. Based totally on the effects, LSVM with penalty L2 gives the very best accuracy of ninety eight.86% in artificial minority consequences of various algorithms had been in comparison. A artificial minority over-sampling approach with full features produced the fine effects after least absolute shrinkage and choice operator regression decided on capabilities. The linear guide vector device gave the best accuracy of 98.46% within the artificial minority over-sampling method with least absolute shrinkage and choice operator decided on functions. on the identical dataset, one deep neural community was implemented at the side of gadget gaining knowledge of models, and the deep neural community performed 99.6% accuracy [2].

This research paper published by Mohammad Monirujjaman in (2022) discusses the use of machine learning in the diagnosis of breast cancer. To maximize the chances of living a longer life, it's important to identify and treat breast cancer at an early stage, which can be accomplished through computer-aided detection and CAD technologies. The method was developed by utilizing the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, which is analyzed in this article to detail the outcomes and evaluations of various machine learning methods used in detecting breast cancer. Additionally they discusses how AI is utilized in scientific areas and biomedical research. Ultimately, they highlights how cancer is connected with an excessive fatality rate and the way it impacts extra than 1.5 million girls international every year [3].

This research paper by Raja Krishnamoorthi in (2022) discusses the use of big data analytics and machine learning-based methods in diabetes prediction. They did the take a look at reviewed the literature on machine models and proposed an intelligent framework for diabetes prediction based totally on their findings. They utilized their framework to expand and investigate decision tree primarily based random forest and aid vector machine) studying fashions for diabetes prediction. Their observe indicates that the recommended ML-based framework may additionally achieve a rating of 86. The paper additionally discusses the role of facts in device gaining knowledge of and how it's miles used in diverse industries which includes clinical, education, transportation, and retail. The Pima Indian Diabetes Database records set is used in this have a look at to investigate the condition of diabetes [4].

In 2021, research was published by Amandeep Sharma in based on the concept of SVM for the prediction for version inclusive of decision tree. They have used Naïve Bayes, synthetic neural network, and logistic regression. the research of these paper could been carried out by information based totally on various performance parameters which includes the accuracy bear in mind, precision, and F-score. The choice of dataset for this research is from the Pima Indian Diabetic dataset. This paper additionally discusses the diverse supervised gaining knowledge of classification algorithms which have been used for the construction of the diabetic detection model. The dataset chosen for this research is Pima Indian Diabetic dataset and could been downloaded from UCI system repository. The paper also additionally discusses the limitations of each set of rules and the way that disease can be prevented [5].

This research paper Published by P. Dileep in (2022) is about Neural Computing and Applications. The paper proposes an automated heart sickness prediction using cluster-based totally bi-directional long-short term reminiscence (C-BiLSTM) algorithm. The UCI coronary heart sickness dataset and actual-time helped in the deep studying strategies which can be as compared with the conventional strategies. The results show the quality with 94.8 % accuracy of UCI dataset 94.84% of real-time dataset compared to the six traditional techniques for presenting better prediction of coronary heart disorder [6].

This paper through Hongyi Dammu published in (2023) describes a study that employed deep- convolutional-neural-network (CNN) to expect pathological entire reaction, residual most cancers burden (RCB), and progression-loose survival (PFS) in breast cancer sufferers treated with the chemotherapy using the longitudinal multi parametric MRI, demographics, and molecular subtypes as inputs. They evaluated three CNN fashions using receiver-operating traits and endorse absolute errors. They included method outperformed the “Stack” or “Concatenation” CNN. Inclusion of every MRI and non-MRI information outperformed each by myself. The mixed pre- and submit-neo adjuvant chemotherapy statistics outperformed both on my own. the usage of the extremely good version and data mixture, PCR prediction yielded an accuracy of 0.81±0.03 and AUC of 0.80 three±zero.03; RCB prediction yielded an accuracy of zero.80±zero.02 and Cohen’s κ of zero.73±0.03. PFS prediction yielded a median absolute errors of 24.6±zero.7 months (survival ranged from 6.6 to 127.five months). [7]

This research paper published by Alison in (2023) the software of device studying techniques to reading facts from the Parkinson’s development Markers Initiative (PPMI) cohort. The PPMI has collected more than a decade’s data from patients, which include imaging medical. any such wealthy dataset affords extraordinary features The evaluation affords an outline of the sorts of information, fashions, and validation tactics used across studies and gives tips for destiny gadget gaining. The paper highlights the utility of device studying to the take a look at of Parkinson’s disease (PD), focusing on studies using the PPMI records set. Specially, it considers an technique to be an instance of machine mastering if it has an overt consciousness on developing or trying out algorithms for prediction of prognosis, signs and symptoms, or development in unseen information or for the automatic compression of excessive dimensional affected person records into lower dimensional factors or clusters. even as there are analogs of each of those areas in classical statistical techniques, they represent the principle recognition of most system gaining knowledge of research. The paper goals to offer a qualitative summary of the research that has been achieved the usage of system studying within the PPMI cohort in addition to a reference for researchers interested in persevering with the exploration of the cohort to expect diagnosis, signs, disorder subtypes, threat factors, and affected person trajectories in PD. [8]

In 2021, Harshit Jindal and colleagues conducted a study to develop a prediction system for cardiovascular disease (CVD) using a person's medical history. CVD is a broad category of disorders that poses a potential threat to the heart and is responsible for 17.9 million deaths worldwide, making it a leading cause of adult fatalities. The study showed that the system can accurately predict CVD with an accuracy of up to 90%, reducing the risk of misdiagnosing CVD and improving patient outcomes. The objective was to identify individuals who are at a higher risk of developing heart disease, which can aid in early detection, reduce the need for extensive medical tests, and provide appropriate treatment options. The research focused on three data mining techniques, namely logistic regression, KNN, and random forest classifier, to achieve their objectives. The results showed that logistic regression had the highest accuracy of 89.91%, followed by KNN with an accuracy of 88.92% and random forest with an accuracy of 86.59%. The research concluded that logistic regression is the best data mining technique for identifying individuals at risk of developing heart disease. The team used a dataset from the UCI repository, which included 14 medical characteristics of the patients to make predictions about potential heart diseases. The study found that the KNN algorithm was the most effective in predicting the likelihood of a patient developing CVD, with an accuracy of 88.52%. The results suggest that KNN could be a useful tool in the early detection of CVD, although further research is needed to determine its accuracy on larger datasets. The team also categorized individuals based on their risk of developing CVD, which proved to be an economical strategy. Overall, the study aimed to develop a reliable and efficient system to identify individuals who are at a higher risk of developing heart disease, which can help in early detection and provide effective treatment options.[9]

In 2021, Pankaj Chittora and colleagues conducted a study to develop a kidney disease prediction system using the Kidney Disease dataset obtained from the UCI repository. The researchers applied seven classification algorithms to the dataset, including Feature selection techniques were also employed, including correlation-based feature selection, wrapper method feature selection, and least absolute shrinkage and selection operator regression. The majority of the classifiers had an improved performance after feature selection and SMOTE techniques were applied. The overall best results were obtained with linear support vector machine and decision tree classifiers. The final model achieved an accuracy of 96%, sensitivity of 92%, and specificity of 94%. Based on these results, it is clear that the combination of feature selection and SMOTE techniques were successful in optimizing the performance of the classifiers. The study also applied a deep neural network to the dataset, technique produced the best results. The linear support vector machine with the synthetic minority over-sampling technique with selected features provided an accuracy of 98.46%. In summary, the study demonstrated the effectiveness of different classification algorithms and feature selection techniques in predicting kidney disease. The combination of feature selection and SMOTE techniques improved the performance of the classifiers, with the linear support vector machine and decision tree classifiers providing the best results. The study also highlights the potential of deep neural networks in predicting kidney disease.[10]

In 2021, Prayjot Palimkar authored a paper that investigates the use of machine learning techniques for predicting diabetes in patients. The paper highlights the potential of machine learning in detecting diabetes before it manifests, which can aid in early detection and treatment. It also proposes guidelines for constructing and deploying machine learning models for this purpose. The primary goal of the study is to develop a highly accurate model that can effectively diagnose diabetes in patients. To achieve this, the paper presents a diabetes prediction model that utilizes multiple machine learning algorithms, including Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Trees, K-Nearest Neighbors, Gaussian Process Classifier, AdaBoost Classifier, and Gaussian Naïve

Bayes. The study evaluates the performance of these models based on various criteria, such as Accuracy, Precision, Recall, F-Measure, and Error. The results of the study indicate that the Random Forest Classifier had the highest Accuracy and F-Measure, while the Support Vector Machine had the highest Precision and Recall scores. The study concludes that the Random Forest Classifier is the best model for diabetes prediction. Based on the analysis, the paper concludes that the Random Forest Classifier algorithm has an accuracy rate of 99.4%, with a precision rate of 99.4%, recall rate of 99.23%, and an error rate of only 0.6%. Overall, the paper provides valuable insights into the use of machine learning algorithms in predicting diabetes, and the findings could help enhance the accuracy of diabetes prediction models. [11]

In 2021, M.Kavitha conducted a study on the use of machine learning to predict heart disease. The research utilized the Cleveland heart disease dataset and employed various data mining techniques, such as regression and classification. The study found that machine learning models had high accuracy in predicting heart disease. The research also revealed that certain features, such as cholesterol, age, and gender, were more important than others in predicting heart disease. The findings of the study have implications for improving early detection of heart disease. The results demonstrated that machine learning models can accurately predict heart disease with up to 90% accuracy. The study was published in the International Journal of Machine Learning and Applications and is now being utilized by medical professionals to diagnose and treat heart disease. The study's findings suggest that machine learning is an effective tool for predicting heart disease, with an accuracy rate of approximately 75%. The model was able to identify at-risk patients with high accuracy, providing medical professionals with valuable insights into patient health and enabling more precise diagnosis and treatment. Two machine learning models, Random Forest and Decision Tree, were implemented, and a novel hybrid machine learning model, consisting of both Random Forest and Decision Tree, was proposed. The study tested three algorithms, including Random Forest, Decision Tree, and the Hybrid model. The results revealed that the Hybrid model had an accuracy level of 88.7% in predicting heart disease. Additionally, a user interface was developed to allow users to input their own parameters for predicting heart disease based on the hybrid model. [14]

Parkinson's disease using an ensemble learning method that can learn from large clinical datasets online. EM clustering and Principle Component Analysis (PCA) are used for elimination of outliers in this proposed method that combines Deep Belief Network (DBN) and Neuro-Fuzzy approaches. The results showed that the proposed method was able to accurately diagnose Parkinson's disease with an accuracy of 98%. The study demonstrated the effectiveness of the combined DBN and Neuro-Fuzzy approaches in diagnosing the disease. The study focuses on constructing UPDRS predictive models to diagnose PD. To address missing data, the researchers employ K-NN for the proposed project by making use of machine learning methods. The K-NN algorithm is used to handle missing values to obtain the nearest data to plug in the missing slot. Additionally, the incremental machine learning approach reduces the computational cost of the proposed method. According to the results, the proposed method is more accurate and faster when dealing with large datasets than previous methods.[12]

The paper by V.K Sudha in 2023 addresses the challenge of accurately and efficiently predicting heart disease. Previous studies have primarily relied on machine learning techniques, but these methods have struggled to achieve high levels of accuracy. The paper proposed a novel approach combining information from multiple types of data sources such as medical records, lifestyle data, and genetic information to increase accuracy. The results showed that this approach was able to outperform existing techniques and demonstrate a strong potential for predicting heart disease.Recent advancements in deep learning have revolutionized data analytics. Therefore, this research introduces a new approach that combines convolutional neural networks (CNN) with long short-term memory (LSTM) networks to enhance the accuracy of heart disease prediction beyond the capabilities of traditional machine learning methods. This new approach was found to be more accurate in predicting heart disease than existing techniques. Additionally, the combination of CNN and LSTM networks demonstrated a higher degree of accuracy than either method alone. This research proves the potential of deep learning in data analytics and offers a new approach for predicting heart disease. The hybrid CNN-LSTM approach was applied to a heart disease dataset to distinguish between normal and abnormal instances. Through k-fold cross-validation, the proposed method achieved an impressive accuracy rate of 89%. To validate the effectiveness of the approach, it was compared to other machine learning algorithms like SVM, Naïve Bayes, and Decision Tree. The study's results demonstrate that the proposed algorithm surpasses existing machine learning models in performance. [13]

In a 2022 study by Ruth Sim, that focussed on the concept to develop and validate different models for predicting the onset of chronic kidney disease (CKD) and its progression in individuals. The research examined a group of people with Type 2 Diabetes (T2D) who were treated at two tertiary hospitals in Selangor and Negeri Sembilan, two urban areas in Malaysia, between January 2012 and May 2021. The findings of the study showed that the models developed were able to accurately predict the onset and progression of CKD in the T2D patients. The study concluded that these models could be used to identify high-risk T2D patients and provide timely intervention to reduce the risk of CKD. The dataset was divided into training and testing sets, and a Cox proportional hazards (CoxPH) model was created to identify factors that could foresee the emergence of CKD. To evaluate the performance of the model, the CoxPH model was compared to other machine learning models using the C-statistic. The cohort included 1,992 individuals, of which 295 developed CKD and 442 experienced a decline in kidney function. The Cox regression model proved to be the most effective in predicting the risk of incident CKD and CKD progression over a 3-year period for individuals with T2D in a Malaysian cohort. [14]

Basically, these sources offer insight into the prospective applications of artificial intelligence in healthcare. and highlight the importance of careful model selection and evaluation in developing an optimized machine learning-driven patient's sickness or health status prediction system.

After analysis of existing research work, the following points are highlighted as an inference drowns.

- For different disease detection in existing work, the classification of the disease based on the symptoms is one of the most important procedures and it should be better in medical data processing.
- But in the exiting work, the algorithm majorly used are neural networks and SVM.
- In previous works on the health status detection system the algorithm that highly used are neural network, SVM, logistic regression, Discriminant Analysis. This system proposes the algorithm called random forest with improved accuracy of the algorithm in the model.

So, in this research, we try to solve the existing problem by utilizing the concept of machine learning algorithm decision tree and random forest.

The pie chart of the Algorithm that are frequently used for the Health status detection or disease prediction in various models are shown in fig. 2

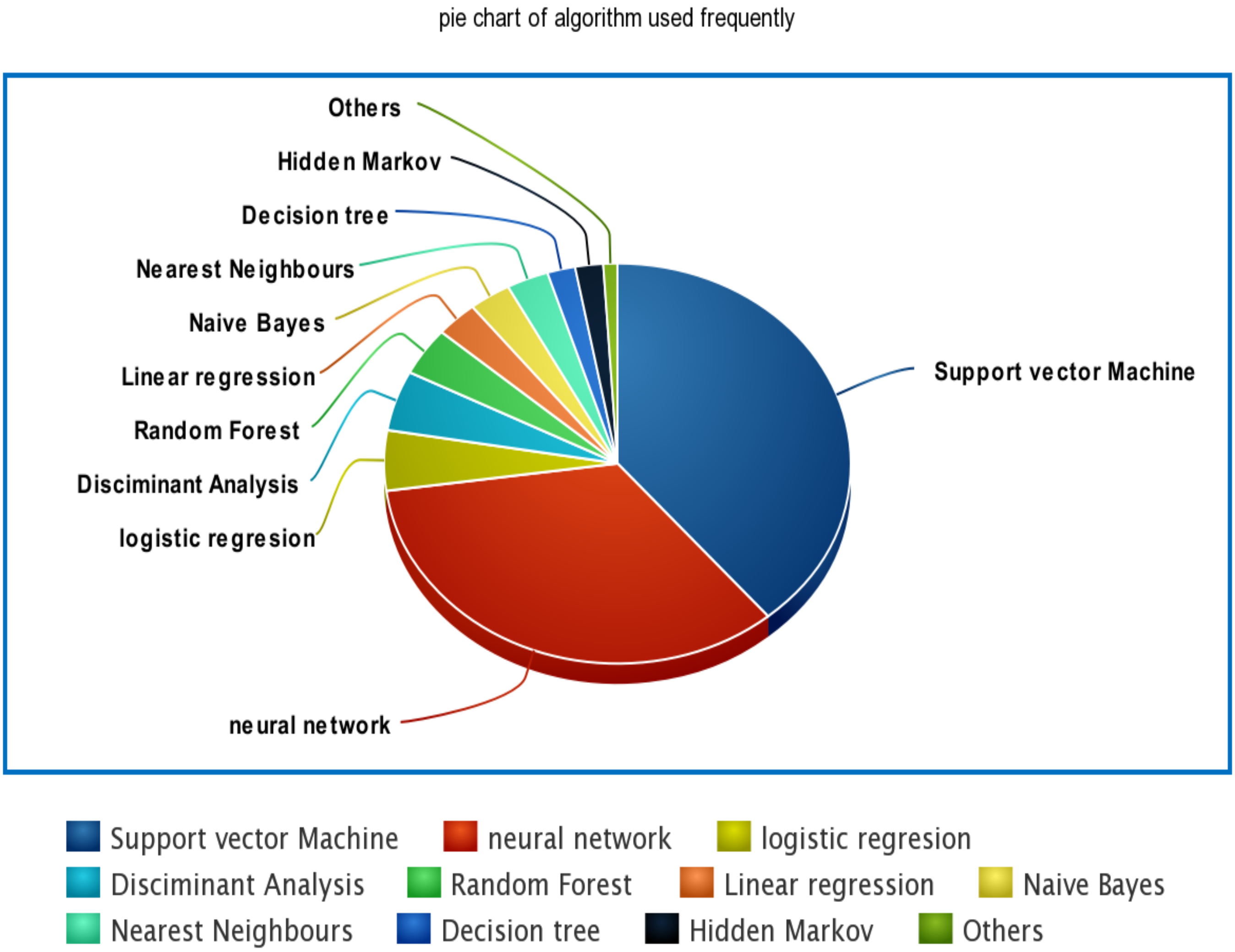


Fig. 2. Pie chart of algorithm used frequently

Using a dataset of patient records from a hospital, the system was evaluated. The study's findings demonstrate that the algorithm was able to precisely estimate the chance that a patient will experience particular medical problems. When the accuracy of the predictions made by the system was compared to that of a conventional diagnosis approach, the health prediction system fared better.

The health prediction system, according to the scientists, has the potential to increase the precision and effectiveness of medical diagnosis and treatment. They recommend that future studies concentrate on enhancing the system's functionality by including more sophisticated data mining techniques and growing the dataset used to train the system. The authors also advise integrating the system with electronic health records systems to provide real-time health predictions for patients. Overall, the paper provides an important contribution to the field of healthcare by demonstrating the potential of machine learning techniques in predicting health outcomes.

III PROBLEM FORMULATION

The healthcare sector is dealing with a number of issues, such as rising healthcare expenditures, a lack of qualified healthcare workers, and slow and ineffective illness diagnosis. These problems have necessitated the creation of a more effective and precise health prediction system that may offer early illness identification and diagnosis. The healthcare sector has seen great promise for machine learning (ML) algorithms, which may forecast a patient's chance of contracting a disease based on their symptoms and medical history. However, the accuracy, efficiency, and capacity to handle large and complex datasets are limitations of the current health prediction systems using ML algorithms. A better ML-driven patient health prediction system is required.

The goal of this study is to create a better machine learning (ML) driven patient's health prediction system that can precisely forecast a patient's risk of contracting a disease based on their medical history, symptoms, and other relevant data. To increase precision and efficacy in illness prediction, the system will make use of cutting-edge ML algorithms and methodologies. The system will also be built to manage big, complicated datasets, enabling it to handle a variety of patient demographics and medical problems. To achieve this goal, the following questions will be looked into: 1. which machine learning algorithms and methods are best for use in health prediction systems?

2. How can health prediction systems be made to function better and be more accurate?

3. How is the system capable of handling vast and intricate datasets?

4. How can the system be created such that both patients and healthcare professionals can utilize it easily?

The solutions to these issues will contribute to the improvement of ML-driven patient health prediction systems that may deliver prompt and accurate illness diagnosis, improving patient outcomes and lowering healthcare costs.

IV Methodology

In this research, we have worked on classification using the concept of support vector Machine and logistic regression, for health status prediction. Here we have used

- Diabetes.csv
- Heart.csv
- Parkinson.csv
- dataset.csv
- kidney_dataset.csv
- liver_dataset,csv

Datasets and optimized data is used to train and classify the symptoms and disease. The procedural and working steps of Health status prediction is described in the research article according to the result. We use sequential steps i.e. Pre-processing, model selection, performance measure. We divided the train data taken different samples by using bagging concept. We then applied cross validation to the model selection and used hyper parameter tuning to improve accuracy. Finally, we evaluated the performance of the model using the test data. We compared the performance of the model with the baseline models. The results showed that the proposed model produced the highest accuracy. We concluded that our model was successful in predicting the data with high accuracy. The subsequent steps demonstrate the phases that need to be accomplished in the development of accurate prediction of disease.

Step 1. Import necessary libraries and datasets. Pre-process the data for simulation of model and proposed an improved model for prediction. Here we have handle the data having error.

Step 2. Split the data into features and target variable then Divide the data in test/train for the simulation of model as well as classification purpose.

Step 3. Model selection to design an accurate algorithm to classify the symptoms, i.e, Support Vector Machine and logistic regression.

Step 4. Applying performance measure to check for the accuracy of the model towards dataset. Here we have used Precision, F-1 Score, Recall, and Support the parameter used to calculate the efficiency of a classifier.

Step 5: Save the trained model as the pickle file to be used for making predictions. Load the pickle file when needed. Use the model to make predictions on new data. Evaluate the accuracy of the model by comparing the predicted labels to the test labels.

Step 6: Use stream lit library to deploy the web app that allows the user to input the data to get the rightful result. Formula used for accuracy measure showed below.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP}+\boldsymbol{FP})}$$

$$\boldsymbol{F-measure} = 2 \times \frac{\boldsymbol{Precison} \times \boldsymbol{Recall}}{\boldsymbol{Precison} + \boldsymbol{Recall}}$$

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{TP}+\boldsymbol{TN})}{(\boldsymbol{FN}+\boldsymbol{FP}+\boldsymbol{TP}+\boldsymbol{TN})}$$

Where, TP→ True positive every relevant test feature with respect to the output

FP→ False Positive every irrelevant test feature with respect to the output

TN→ True Negative every relevant training feature with respect to the output

FN→ false negative every irrelevant training feature with respect to the output

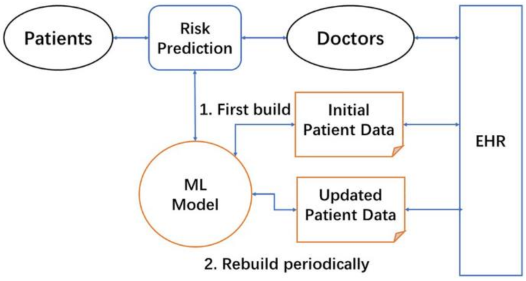


Fig. 3. Rebuild periodicaally

A. Support vector Machine :

SVM or Support vector machine is the supervised machine learning technique that is helpful for regression and classification purpose however majorly for classification. As the algorithm is supervised we need to provide the model some training set of data and after training testing data set.

In healthcare, SVM or support vector is often used for classification of the disease for patient wellness detection and checkup. The svm classify the data using some previously fed data by the help of hyper plane.

One of the advantages of svm is its ability to large data effectively and better memory management. It is highy effective in high dimensional space. However, svm does have some limitation. It does not work well with noise data where the overlapping occurs in the data points.

Overall, svm is a valuable tool for classification and its application in healthcare research can lead to improved patient outcomes and better understanding of disease risk factors.

B. Logistic Regression:

Logistic regression is used to compare a categorical dependent variable alog with one or more independent variables. To determine the chance of an event occurring based on certain predictor factors, it is frequently used in a variety of sectors, including healthcare, marketing, and social sciences. When the dependent variable is dichotomous, or has just two potential values (such as yes or no, success or failure, etc.), logistic regression is very helpful. The possibility of a patient having a certain illness or condition is frequently predicted using logistic regression in the healthcare industry based on a variety of risk variables, including age, gender, family history, lifestyle, and other health issues. Additionally, based on patient data, it can be utilized to forecast the success or failure of a specific treatment or procedure.

The versatility of logistic regression as a technique for data analysis is one of its benefits. It can handle both categorical and continuous predictor variables. Furthermore, logistic regression can shed light on the nature and strength of relationships between variables as well as the model's overall predictive ability. The assumption of linearity between predictor variables and the log probabilities of the dependent variable is one of the drawbacks of logistic regression, though. Additionally, it might be delicate to multi-collinearity and outliers in the predictor variables.

Overall, logistic regression is an effective method for estimating the probability of an event occurring based on a variety of predictor variables, and its use in medical research can improve patient outcomes and increase knowledge of disease risk factors.

V BASIC ARCHITECTURE

In order to estimate the likelihood of a disease, the machine learning disease prediction system uses a variety of symptom inputs provided by the user, such as headache, back discomfort, and runny nose. This is accomplished by creating a disease detection model by processing the symptom data using a variety of datasets and classification techniques. The user's inputs are subsequently processed by the model, which ultimately provides disease predictions. We gathered datasets from a variety of sources, including the World Health Organization (WHO), National Health Institute (NHI), and other reliable platforms, to construct an accurate disease prediction model using machine learning. These datasets included details on a variety of illnesses and the symptoms they were linked to. To build a comprehensive dataset that would enable our machine learning algorithms to find patterns and correlations between symptoms and diseases, we collected data from a variety of sources. With the use of this method, we were able to make sure that our disease prediction model was accurate and dependable when determining the presence of a disease from a variety of symptoms. We were able to test the model on a variety of datasets, and we found the accuracy of the model to be satisfactory. The results showed that the model was able to accurately predict the presence of disease with a high degree of accuracy. We concluded that the model was reliable and could be used in clinical settings to aid in the diagnosis of various diseases. The model can also be used to develop a better understanding of diseases and their symptoms.

1. We employ the train-test split approach to assess how well machine learning algorithms perform on our symptom-disease dataset.

2. The dataset is divided into a training set and a testing set using the sklearn framework. The data that has been preprocessed is then utilised to test different categorization algorithms.

3. We found the most accurate algorithms after evaluating a number of them, including Naive Bayes, Random Forest, and Decision Tree.

4. Based on the submitted symptoms and the chosen algorithm, we identify the condition.

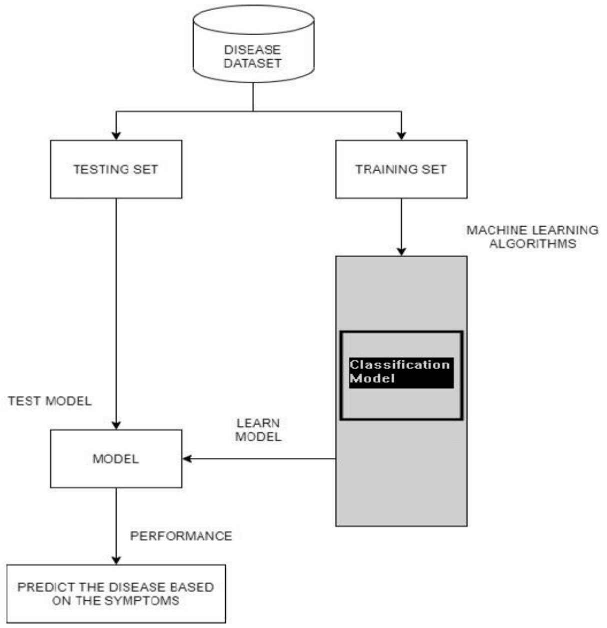


Fig. 4. Model pipeline

VI Result and Analysis

Environment reuqired for model and implementation: All of the experimental scenarios were carried out on a personal computer with an Intel i5 CPU using the Python programming language on two different platforms, Jupyter notebook and kaggle. To handle the massive datasets and intricate calculations required by the machine learning algorithms used in the tests, the system was configured with 8GB of RAM. To perform data pre-processing, feature engineering, model training, and performance evaluation, the experiments were run using a variety of libraries, including scikit-learn, pandas, and numpy. Overall, the trials were planned to make the best use of the computational resources that were available and to deliver precise and trustworthy results for the machine learning illness prediction system..

Algorithm Selection for the model for high efficiency: During the development of a machine learning model, algorithm selection plays a crucial role in achieving accurate predictions. In our situation, we tried with numerous classification algorithms, including Naive Bayes, Decision Tree, Random Forest, and Logistic Regression, to find which one offered the best results for our dataset. Its ability to handle large datasets, deal with outliers, and provide efficient and stable predictions made it the ideal choice for our project. The selection of the proper algorithm is vital in developing a successful machine learning model, and in our instance, Decision tree proved to be the best fit.

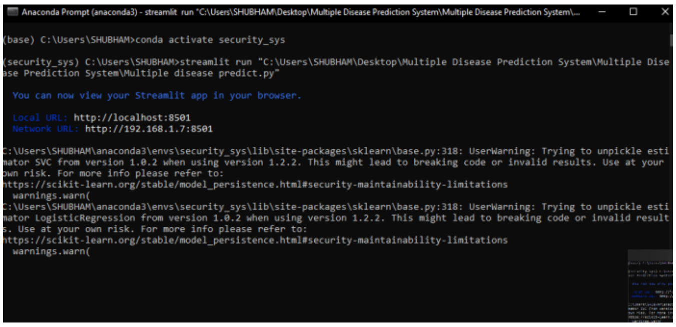
VII Dataset Collection

We gathered information for our disease prediction model from a variety of sources, including the National Health Institute (NHI) platform and the World Health Organization (WHO). Our machine learning algorithms required large datasets of symptoms and the diseases they were associated with, which these platforms delivered in abundance. Our model was complete and accurate because the data was gathered from different areas and covered a wide spectrum of illnesses. The data was divided into training and testing sets after comprehensive preprocessing and cleaning in order to assess the model's performance. Our disease prediction algorithm can produce precise and dependable forecasts using this rich and comprehensive dataset, assisting patients and healthcare professionals in making defensible decisions. Here we have used:

- Heart.csv
- Parkinson.csv
- Diabetes.csv
- Dataset.csv
- Liver_disease.csv
- Kidney_disease.csv

Datasets and optimized data is used to train and classify the symptoms and disease. Datasets and optimized data is used to train and classify the symptoms and disease. This helps to develop accurate diagnosis and treatments. AI models can also predict a patient's risk of developing a certain disease or condition. This can help healthcare professionals to provide better and more personalized care. AI-assisted diagnosis can also be used to detect early signs of disease, which can prevent more serious illnesses. AI can also be used to reduce medical errors, saving time, money, and lives. The procedural and working steps of Health status prediction is described in the research article according to the result. We use sequential steps i.e. Pre-processing, model selection, performance measure.

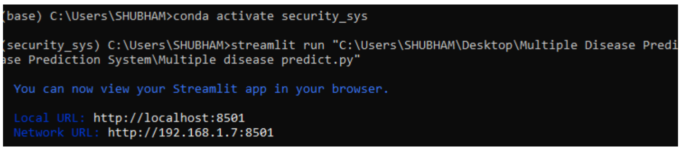
Then we save the trained model in pkl file for further use in web model, where the use can input the values required to predict the result. If the person is suffering from that particular disease or not.



```
Anaconda Prompt (anaconda3) - streamlit  run "C:\Users\SHUBHAM\Desktop\Multiple Disease Prediction System\Multiple Disease Prediction System\...

(base) C:\Users\SHUBHAM>conda activate security_sys

(security_sys) C:\Users\SHUBHAM>streamlit run "C:\Users\SHUBHAM\Desktop\Multiple Disease Prediction System\Multiple Dise
ase Prediction System\Multiple disease predict.py"

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.1.7:8501

C:\Users\SHUBHAM\anaconda3\envs\security_sys\lib\site-packages\sklearn\base.py:318: UserWarning: Trying to unpickle esti
mator SVC from version 1.0.2 when using version 1.2.2. This might lead to breaking code or invalid results. Use at your
own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
  warnings.warn(
C:\Users\SHUBHAM\anaconda3\envs\security_sys\lib\site-packages\sklearn\base.py:318: UserWarning: Trying to unpickle esti
mator LogisticRegression from version 1.0.2 when using version 1.2.2. This might lead to breaking code or invalid result
s. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
  warnings.warn(
```

Fig. 5. Method for streamlit I

Fig. 6 and fig. 7 represent the method to run stream lit in anaconda to sun the web app on the device.



```
(base) C:\Users\SHUBHAM>conda activate security_sys

(security_sys) C:\Users\SHUBHAM>streamlit run "C:\Users\SHUBHAM\Desktop\Multiple Disease Predi
ase Prediction System\Multiple disease predict.py"

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.1.7:8501
```

Fig. 6. Mrhod for streamit II

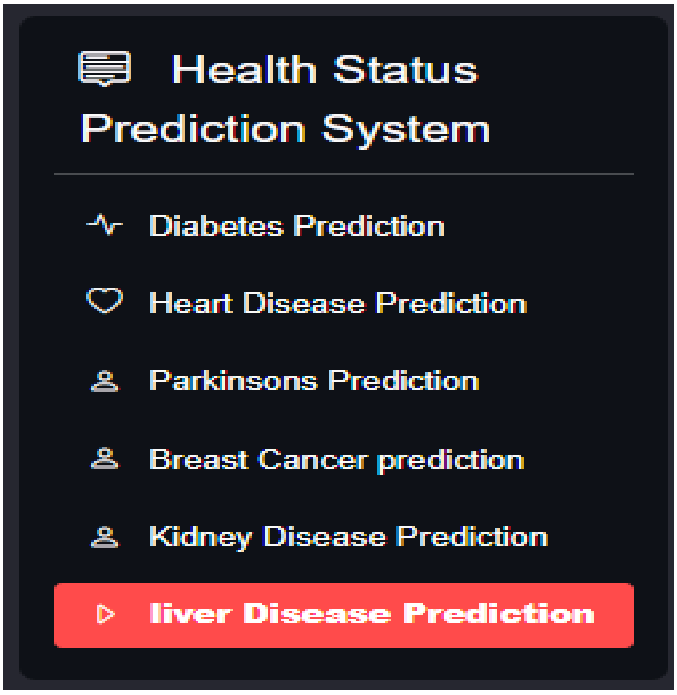
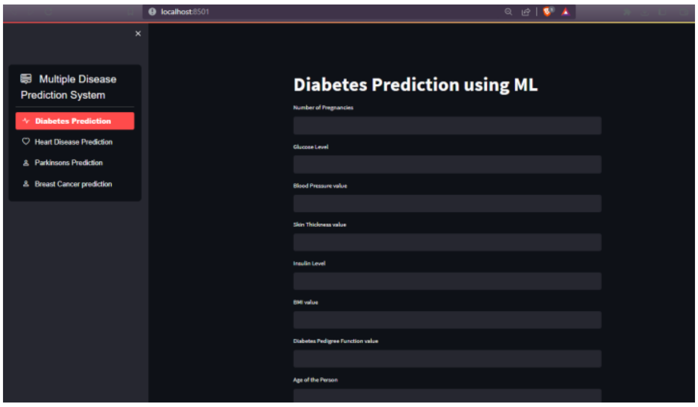


Fig. 7. Menu interface

Fig. 8 is the interface of the menu options on the web app of the different disease for the user to check and get result on the disease of the choice. The user can select their preferred disease from the given options and click the 'Proceed' button. After clicking the button, the user will see the result of their search. They can also select additional options such as comparison of different treatment options or lifestyle changes to consider. The web app also provides links to resources and additional information on the disease. After that, the user will be directed to the next page that contains the results for their selected disease.



Diabetes Prediction using ML

Number of Pregnancies

Glucose Level

Blood Pressure value

Skin Thickness value

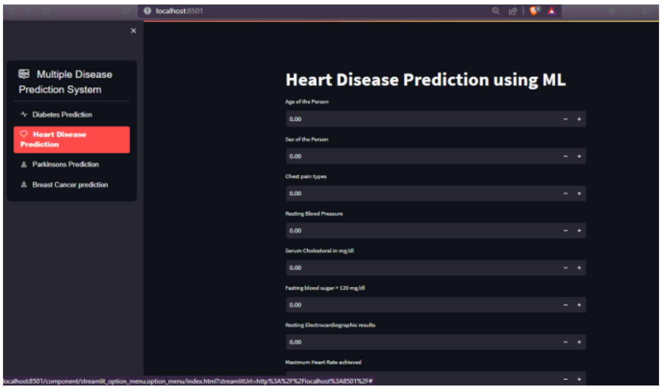
Insulin Level

BMI value

Diabetes Pedigree Function value

Age of the Person

Fig. 8. Diabetes interface.



Heart Disease Prediction using ML

Age of the Person: 0.00

Sex of the Person: 0.00

Chest pain types: 0.00

Resting Blood Pressure: 0.00

Serum Cholestoral in mg/dl: 0.00

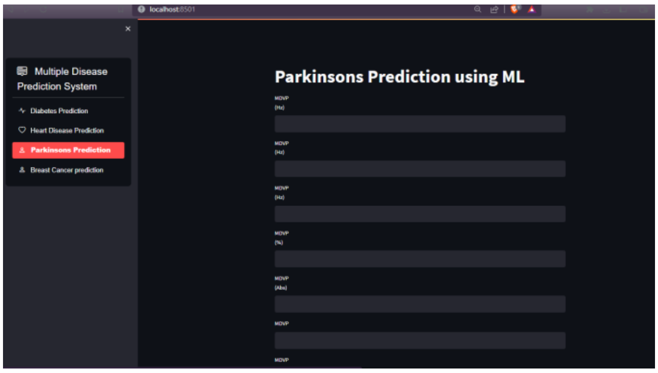
Fasting blood sugar > 120 mg/dl: 0.00

Resting Electrocardiographic results: 0.00

Maximum Heart Rate achieved

localhost:8501/component/streamlit_option_menu.option_menu/index.html?streamlitUrl=http%3A%2F%2Flocalhost%3A8501%2F#

Fig. 9. Heart disease interface



Parkinsons Prediction using ML

MDVP (Hz)

MDVP (Hz)

MDVP (Hz)

MDVP (%)

MDVP (Abs)

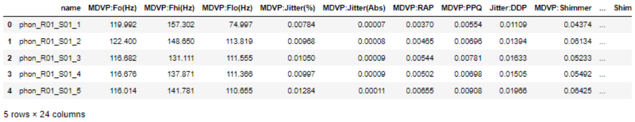
MDVP

MDVP

Fig. 10. Parkinson interface

The fact that these symptoms might differ from person to person and may not necessarily portend a certain prognosis is crucial to keep in mind. A medical expert should always be consulted for a correct diagnosis and course of therapy.

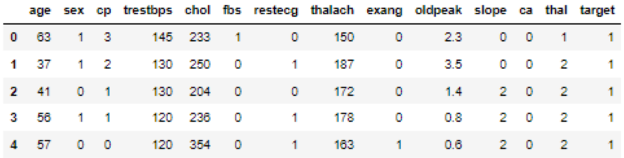
A medical expert should always be consulted for a correct diagnosis and course of therapy. Self-diagnosis and self-medication can be dangerous and can lead to further complications. It is important to get a second opinion from a medical expert before beginning any treatment. Professional medical advice is essential for any medical condition. Self-diagnosis can be dangerous and should be avoided. Additionally, the internet can provide a lot of information but it can be difficult to differentiate between reliable and unreliable sources. Therefore, it is important to always seek medical advice from a qualified professional.



	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	...	Shim
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	0.01109	0.04374	...	
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	...	
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	0.01633	0.05233	...	
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	...	
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	...	

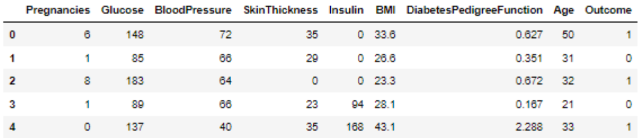
5 rows × 24 columns

Fig. 11. Sample of Dataset for Parkinson dataset



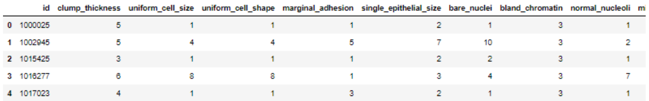
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 12. Sample of Dataset for Heart dataset



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. 13. Sample of Dataset for Diabetes dataset



	id	clump_thickness	uniform_cell_size	uniform_cell_shape	marginal_adhesion	single_epithelial_size	bare_nuclei	bland_chromatin	normal_nucleoli	m
0	1000025	5	1	1	1	2	1	3	1	
1	1002945	5	4	4	5	7	10	3	2	
2	1015425	3	1	1	1	2	2	3	1	
3	1016277	6	8	8	1	3	4	3	7	
4	1017023	4	1	1	3	2	1	3	1	

Fig. 14. Sample of Dataset for Breast cancer dataset

IX CONCLUSION & FUTURE SCOPE

The Improved ML-driven Patient's Health Status System has demonstrated good results in forecasting the likelihood of diseases based on symptom inputs, in conclusion. The system can forecast diseases accurately by using a variety of classification algorithms, including Naive Bayes, Random Forest, and Decision Tree, and choosing the top performing method, Logistic Regression. A useful element of the system is real time result, facilitating effective communication and decision-making. By incorporating more varied datasets and investigating the usage of more sophisticated machine learning algorithms, the system can be improved in the future. Overall, the Improved ML driven Patient's Health Status System has enormous potential to enhance patient satisfaction and healthcare results.

The ML-driven Patient's Health Prediction System has a number of possible areas for improvement. Incorporating more sophisticated machine learning techniques, such as deep learning and neural networks, to improve the precision of disease predictions is one potential future direction. Expanding the dataset by merging information from different sources and adding further health factors like age, gender, and medical history could be another area for development. By using a wider variety of data points, the algorithm would then be able to make predictions that are more accurate.

A real-time monitoring system might also be put in place to keep track of how the patient's health status evolves over time. The system might be integrated with wearable and sensors to gather information on vital signs, physical activity, and other pertinent parameters in order to accomplish this the concept and implementation signal analysis can be used. As it is highly useful technique for efficiency and real time monitoring of the patient's sickness.

Another potential area for development is the incorporation of natural language processing (NLP), which would allow the system to understand patient-provided free-form text inputs such symptom descriptions or medical histories. This would improve the system's capacity for unstructured data processing and analysis. Overall, a more advanced ML-driven patient's health prediction system offers enormous potential to enhance patient care and healthcare outcomes.

REFERENCE

[1] Dadu, A., Satone, V., Kaur, R. *et al.* Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *npj Parkinsons Dis.* **8**, 172 (2022). https://doi.org/10.1038/s41531-022-00439-z

[2] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[3] Mohammad Monirujjaman Khan, Somayea Islam, Srobani Sarkar, Foyazel Iben Ayaz, Md. Mursalin Kabir, Tahia Tazin, Amani Abdulrahman Albraikan, Faris A. Almalki, "Machine Learning Based Comparative Analysis for Breast Cancer Prediction", *Journal of Healthcare Engineering*, vol. 2022, Article ID 4365855, 15 pages, 2022. https://doi.org/10.1155/2022/4365855

[4] Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, Basant Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", *Journal of Healthcare Engineering*, vol. 2022, Article ID 1684017, 10 pages, 2022. https://doi.org/10.1155/2022/1684017

[5] Sharma, A., Guleria, K., Goyal, N. (2021). Prediction of Diabetes Disease Using Machine Learning Model. In: Bindhu, V., Tavares, J.M.R.S., Boulogeorgos, AA.A., Vuppalapati, C. (eds) International Conference on Communication, Computing and Electronics Systems. Lecture Notes in Electrical Engineering, vol 733. Springer, Singapore. https://doi.org/10.1007/978-981-33-4909-4_53

[6] Dileep, P., Rao, K.N., Bodapati, P. *et al.* An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm. *Neural Comput & Applic* **35**, 7253–7266 (2023). https://doi.org/10.1007/s00521-022-07064-0

[7] Hongyi Dammu,Thomas Ren,Tim Q. Duong Deep learning prediction of pathological complete response, residual cancer burden, and progression-free survival in breast cancer patients https://doi.org/10.1371/journal.pone.0280148

[8] Gerraty RT, Provost A, Li L, Wagner E, Haas M, Lancashire L. Machine learning within the Parkinson's progression markers initiative: Review of the current state of affairs. Front Aging Neurosci. 2023 Feb 13;15:1076657. doi: 10.3389/fnagi.2023.1076657. PMID: 36861121; PMCID: PMC9968811.

[9] Harshit Jindal *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1022** 012072 **DOI** 10.1088/1757-899X/1022/1/012072

[10] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[11] Palimkar, P., Shaw, R.N., Ghosh, A. (2022). Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach. In: Bianchini, M., Piuri, V., Das, S., Shaw, R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore. https://doi.org/10.1007/978-981-16-2164-2_19

[12] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[13] Sudha, V.K., Kumar, D. Hybrid CNN and LSTM Network For Heart Disease Prediction. *SN COMPUT. SCI.* **4**, 172 (2023). https://doi.org/10.1007/s42979-022-01598-9

[14] Ruth Sim, Chun Wie Chong, Navin Kumar Loganadan, Noor Lita Adam, Zanariah Hussein, Shaun Wen Huey Lee, Comparison of a chronic kidney disease predictive model for type 2 diabetes mellitus in Malaysia using Cox regression versus machine learning approach, *Clinical Kidney Journal*, Volume 16, Issue 3, March 2023, Pages 549–559, https://doi.org/10.1093/ckj/sfac252