

Assistance Tool for Prediction and Monitoring Various Diseases based on Machine Learning

Saurbh Singh Jamwal¹, Vishesh Chaudhary², Ashish Shrivastava³

^{1,2}Students, School of Computer Science and Engineering, Lovely Professional University, Phagwara, India

³Assistant Professor, School of Computer Science and Engineering, Lovely Professional University, Phagwara, India

Abstract - We are living in the era where everybody is busy in their duties. People do not have enough time to care about their own body and as a result most of the time people just ignore the signs that our body shows as a result of something that may be problematic in future. In order to make the process of threat recognition easier, a medium is required which can warn people about the diseases that might be dormant in their body. It will also reduce the amount of time required to perform the diagnosis as it can be done anywhere using compatible system.

Key Words: Naïve Bayes, Random Forest, K-Nearest Neighbour, Python, Health Prediction, Machine Learning.

1. INTRODUCTION

Our project will be predicting disease and severity based on symptoms and data provided by the user. To apply prediction on user symptoms we will be looking at how different data present at our disposal can help us in predicting what the possible disease is. We compared various algorithms to see which one performed best. Further categorization was required to make both specialized as well as generalized treatment possible with ease. The specialized section can later be expanded by adding data for more body parts. This interface will save time, which is an essential commodity no matter what times we live in. The severity of symptoms can also be considered to predict how dangerous the symptoms are, which will indicate what degree of care needs to be taken care.

2. LITERATURE REVIEW

In the paper "Disease Prediction System" *Sarthak Khurana et al. [1]* it was discussed how the symptoms of the disease can be used to evaluate the disease that a person is suffering. Machine learning techniques like Logical Regression, Decision Trees (DT), Random Forest and Naïve Bayes techniques have been used in order to get the required result. The user was given liberty to choose what technique one wants to use for the prediction but there were some limitations such as lack of information about the disease predicted by the system.

In "Virtual Doctor" *Divyansh Tiwari et al. [2]* discussed how data mining can be used for finding unknown values in large data. It aimed for a mechanized diagnosis system that supported the medical practitioner to make a good decision in treatment and disease. The study was aimed mainly for health concerned people and for those who wants to have their own doctor. It was an interactive service for users who wants to know about what health issues they are going through as per the symptoms.

In the paper "A study on data mining prediction techniques in healthcare sector" *Megha Rathi and Vikas Pareek [3]*, information discovery method Knowledge Discovery in Databases (KDD) was used by the authors as the method of adjusting the low-level data into high-level knowledge. Hence, KDD was used for the non-trivial removal of implicit, antecedent, unknown and doubtless helpful data from information in databases. The method consisted of the subsequent steps including information cleansing, information integration, information choice, and information transformation. In healthcare data processing prediction, the supported data processing techniques have used machine learning algorithms such as neural network, Bayesian Classifiers, Decision trees and Support Vector Machine.

In the paper "A Medical Document Classification System for Heart Disease Diagnosis Using Naïve Bayesian Classifier" *D.J.S. Sako et al. [4]* authors have used machine learning algorithm of Naïve Bayes for generating heart diseases prediction based on factors such as age, sex, resting blood pressure, etc. The research made use of the fact that nowadays there is enormous amount of medical data being available in electronic form so we can predict based on previous medical records of patients. The method is a simple probabilistic classifier based on applying Bayes theorem. The features from training datasets have been used to formulate probabilistic models that helped in predicting new data on those features. The Waikato Environment for Knowledge Analysis (WEKA) tool which uses JAVA machine learning methods has been used.

In the paper "Heart Disease Prediction Using Machine learning and Data Mining Technique" *Jaymin Patel et al. [5]* have discussed how heart diseases have become a big threat. They discussed the fact that in order to reduce the number of deaths from heart diseases there must be a quick and efficient detection technique. Decision Tree is one of the effective data mining methods used. This research compared different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis using WEKA tool.

In the paper “Data Mining for Medical System” **Muhamad Hariz Muhamad Adnan et al. [6]** have used multiple models and calculated the accuracy of these models, some Machine Learning models have been used were Fine, Medium and Coarse Decision trees, Gaussian Naïve Bayes, Kernel Naïve Bayes, Fine, Medium and Coarse K-Nearest Neighbor (KNN), Weighted KNN, and Subspace KNN and noted that there was a huge difference in the accuracy of these algorithms.

In the paper “Heart disease prediction using machine learning techniques” **VV Ramalingam [7]** has discussed how supervised learning algorithms such as Support Vector Machines, KNN, Naïve Bayes, DT, Random Forest and ensemble models were found very popular among the researchers. The research also provided evidence that these algorithms were much better than others in terms of the results and hence proved more appropriate to be used in data mining projects.

In the paper “Disease Prediction using Machine Learning” **Kedar Pingale et al. [8]** authors have used Naïve Bayes algorithm, for clustering KNN algorithm, final output had been calculated in the form of 0 or 1 for which Logistic tree is used. The interface accepted the structured and textual type of data as input to the machine learning mode. The end calculations used Logistic Regression and the system was used for predicting chronic diseases based on symptoms.

In the paper “Prediction of Probability of Disease Based on Symptoms Using Machine Learning Algorithm” **Harini DK and Natesh M [9]** have combined both the structured and unstructured data in healthcare field to assess the risk of disease. For S-data, they used three conventional machine learning algorithms, i.e., Naïve Bayes, KNN, and DT to predict the risk of disease. For T-data, CNN-based Unimodal Disease Risk Prediction algorithm was proposed to predict the risk of disease.

In the paper “Disease Prediction Based on Symptoms Using Classification Algorithm” **Sneha R et al. [10]** have used Decision Trees to predict the disease patient is suffering based on the symptoms. Decision tree classifier is trained on the dataset which will come up with the right questions to ask. Internally it builds a decision tree trying to separate diseases from each other at every node. Decision trees traverse from the root to one of the leaves based on decisions made by them.

3. PROPOSED WORK

The main goal of this interface is to predict the health condition of a person based on the symptoms/parameters related to the user. The whole prediction is divided into two parts Generalized, basic disease prediction based on the symptoms the user is experiencing, and Specialized, to predict health condition focusing on organ/ body part.

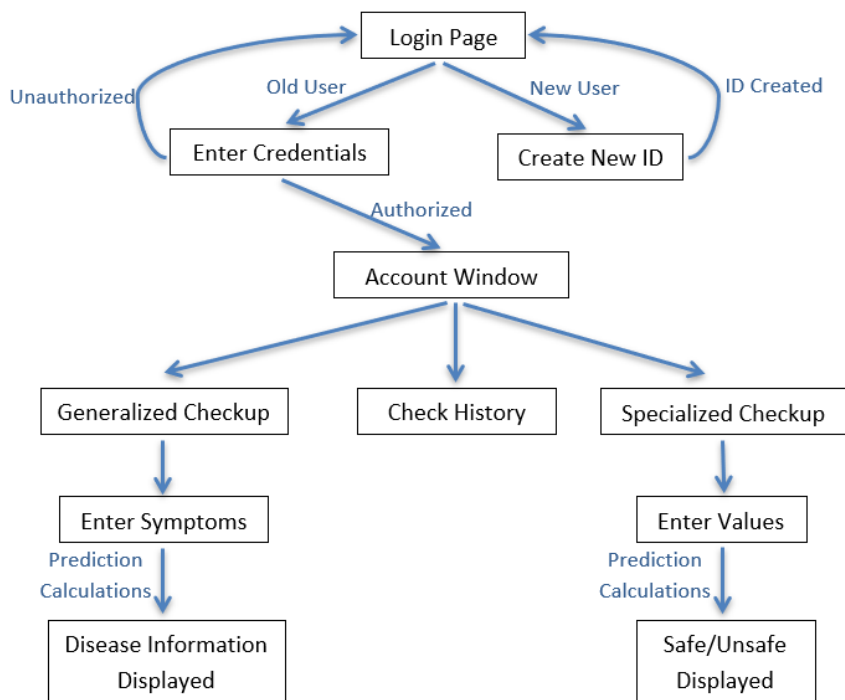


Fig -1: State Diagram of Proposed Work

As the author mentioned in Paper [1], symptoms can be basis of predicting the disease a person is having. In Paper [5] it was shown how some basic parameters can be used to predict presence of heart disease. The algorithms used in the prediction are: **Random Forest, KNN and Naïve Bayes.**

3.1 Methodology

The objective of this model is to simplify the process of disease prediction using the large amounts of patient details available from previous checkups.

3.1.1 Data Collection

Table -1: Dataset for Heart Patients

NAME	TYPE	DESCRIPTION
Age	Continuous	Age in years
Sex	Discrete	0=female, 1=male
Chest Pain Type	Discrete	1= typical Angina, 2= atypical angina, 3= non-Angina pain, 4= Asymptomatic
Resting BP	Continuous	Resting Blood Pressure
Cholesterol	Continuous	Serum Cholesterol Mg/dL
FBS	Discrete	Fasting blood sugar>120 mg/dl: 1=true 0=False
RER	Discrete	Resting Electrocardiographic Results: values 0,1 and 2
Max Heart Rate	Continuous	Maximum heart rate achieved
Induced Angina	Discrete	Exercise Induced Angina 1=Present, 0= Absent
Old Peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	Slope of peak exercise segment 1= up sloping, 2= flat, 3= down sloping
NMV	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect

The dataset for Heart Disease has attributes as shown in **Table 1**. The dataset is the same as used in Paper [5].

Table -2: Dataset for Symptoms and Diseases

NAME	TYPE	DESCRIPTION
Disease	Discrete	Name of disease
Disease ID	Continuous	Uniquely identifies a disease using an integral value
Description	Discrete	Gives description of the disease in form of string value
Precautions	Discrete	Gives 3 precautions that can be taken to avoid the disease
Symptoms	Discrete	Gives name of symptoms for the disease
Symptom ID	Continuous	Uniquely identifies a symptom using integral value

Symptom Severity	Discrete	Tells about the severity of symptoms, value ranges from 1 to 7 as low to high
------------------	----------	---

The dataset used for generalized disease prediction using symptoms has following attributes as shown in **Table 2**. In Paper [3] and Paper [7] the authors discussed why data cleaning is required in prediction models.

3.1.2 Data Preparation

It is essential to treat the outliers and missing values before using the dataset. For this first data exploration was done. Then Tableau Prep Builder was used to achieve the task of data preparation. Inconsistent data was treated using it.

3.1.3 Model Training

The algorithms used in the project are mentioned below. In order to decide which technique is to be used for the data prediction, first it is checked as to which of the above methods are giving better accuracy when predicting values for training dataset. Authors in Paper [1] discussed how many algorithms can be compared based on testing and training dataset performances. For this first the accuracy for each of the algorithms is checked on both specialized and generalized datasets. As shown in Paper [7], following algorithms proved to work better for such data.

Random Forest:

Random Forest is one great algorithm for training early in the model development process and it is very hard to create a “bad” Random Forest. The above fact is additionally an excellent alternative, when one has to save the overall time. Paper [10] showed how Random Forest makes decision and choses which way to go. It provides a credible indicator of the importance it assigns to your options. Random Forests are terribly difficult to ram down in terms of performance. They can easily handle lots of various feature varieties, such as binary, categorical and numerical. It was shown in Paper [5] that lot of variations of Random Forest algorithm are very accurate.

Naïve Bayes:

It is a classification technique based on the mathematical Bayes Equation which has conditions of event happened already, also called as conditional probability. Paper [9] indicated how Naïve Bayes can perform predictions with good accuracies for such datasets, as it makes assumptions of independence among predictors. The main feature of this family of algorithms is that every pair of features being classified is independent of each other. It requires less training data and can handle both continuous and discrete values. It is also highly scalable with options to change number of predictors and data points. Overall, it is very fast and hence useful in real time predictions.

KNN:

The K-Nearest Neighbor algorithm is a supervised machine learning algorithm. It can be used for both classification and regression purposes. It is comparatively easier to implement and understand, but comes with a drawback of being slower as the amount of data is incremented. Its uses and functionalities have been illustrated in Paper [9] along with accuracy. It then plots the training data points into a multidimensional plane. After the predictions are made, the neighbors of point to be predicted are looked at and based on the value of K, polling is done between these many (K) members that are closest to the point of interest.

3.2 Result Discussion

3.2.1 Accuracy of the Results for the Specialized Heart Checking Prediction

The testing dataset used had 30 records whose results were to be predicted and then gets compared with what the actual answer should have been. The accuracies for all three algorithms are as shown by plots in **Chart 1**, **Chart 2** and **Chart 3**. Along with these charts, confusion matrices are also added in **Figure 3**, **Figure 4** and **Figure 5**.

A. Random Forest:

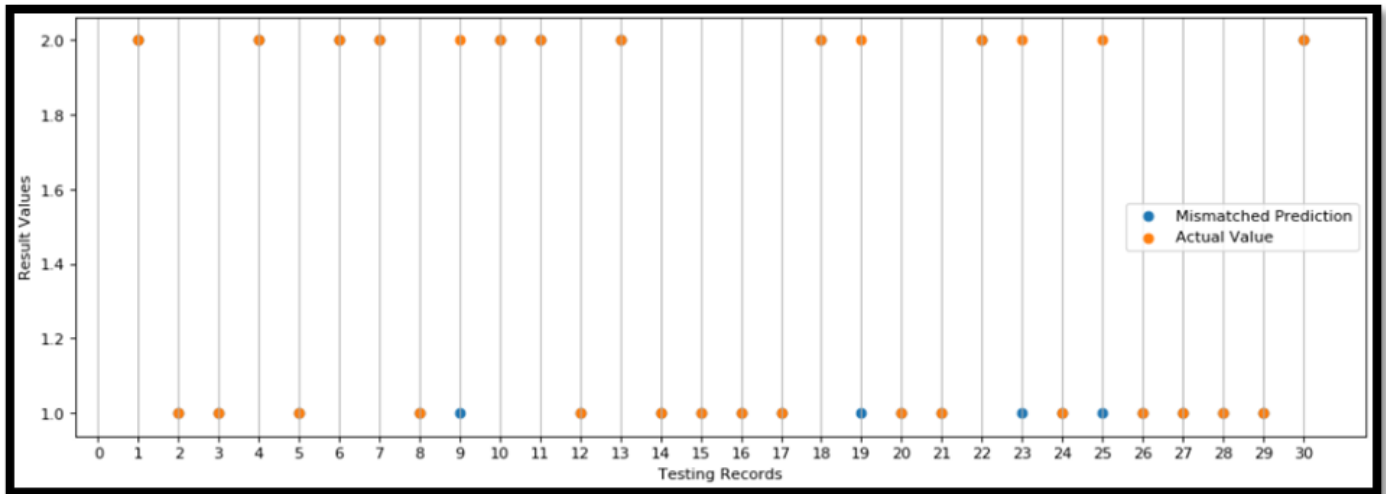


Chart -1: Scatter Plot for Heart Prediction using Random Forest

As shown in **Chart 1**, the testing dataset has 30 records. Using Random Forest as our model, we received 4 mismatched values out of 30 total predictions.

Confusion Matrix, also called Error Matrix, is table layout which can visualize performance of an algorithm and is shown in **Figure 2**. It is an important indicator of model accuracy as demonstrated in Paper [1].

	0	1
0	TN	FP
1	FN	TP

Where:

TN is True Negative

FP is False Positive

FN is False Negative

TP is True Positive

Fig -2: Confusion Matrix Structure

This can be drawn for predictions where there are only two possible values, such as 0 or 1. It can be any binary values. Confusion Matrix for Random Forest is as shown in **Fig 3**.

	0	1
0	10	0
1	4	16

$$\text{Accuracy} = (TP + TN) / \text{Total number of records}$$

$$= (16 + 10) / 30 = 26 / 30 = 86.67\%$$

Fig -3: Confusion Matrix for Random Forest

B. Naïve Bayes:

As shown in the **Chart 2**, there are total 5 mismatched predictions, 1 is false positive and 4 are false negatives.

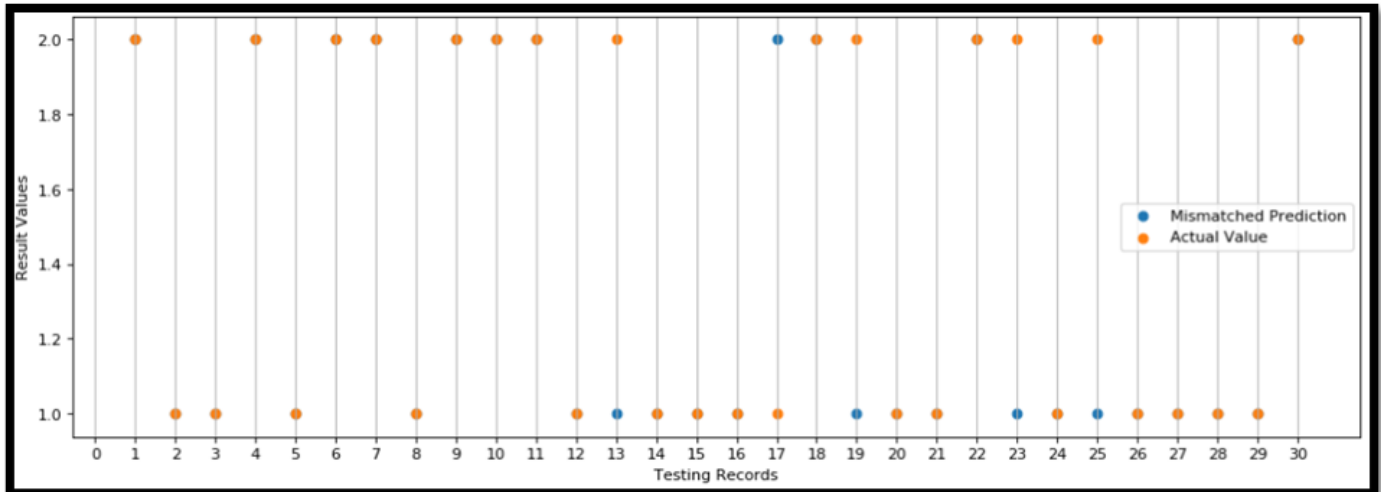


Chart -2: Scatter Plot for Heart Prediction using Naïve Bayes

In total there are 5 mismatched values out of 30 predictions made. Along with this, Confusion Matrix for Naïve Bayes is shown in **Fig 4** and with it the accuracy calculations are also done in same figure.

	0	1
0	10	1
1	4	15

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Total number of records}$$

$$= (15 + 10) / 30 = 25 / 30 = \mathbf{83.33\%}$$

Fig -4: Confusion Matrix for Naïve Bayes

C. K-Nearest Neighbor

As shown in **Chart 3**, the performance of KNN is the least out of all the algorithms used. There are 5 False Positives and 5 False Negatives. So, there are 10 mismatched results out of 30 predicted values. The Confusion Matrix for Euclidean KNN is shown in **Fig 5**.

	0	1
0	9	5
1	5	11

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Total number of records}$$

$$= (11 + 9) / 30 = 20 / 30 = \mathbf{66.67\%}$$

Fig -5: Confusion Matrix for KNN

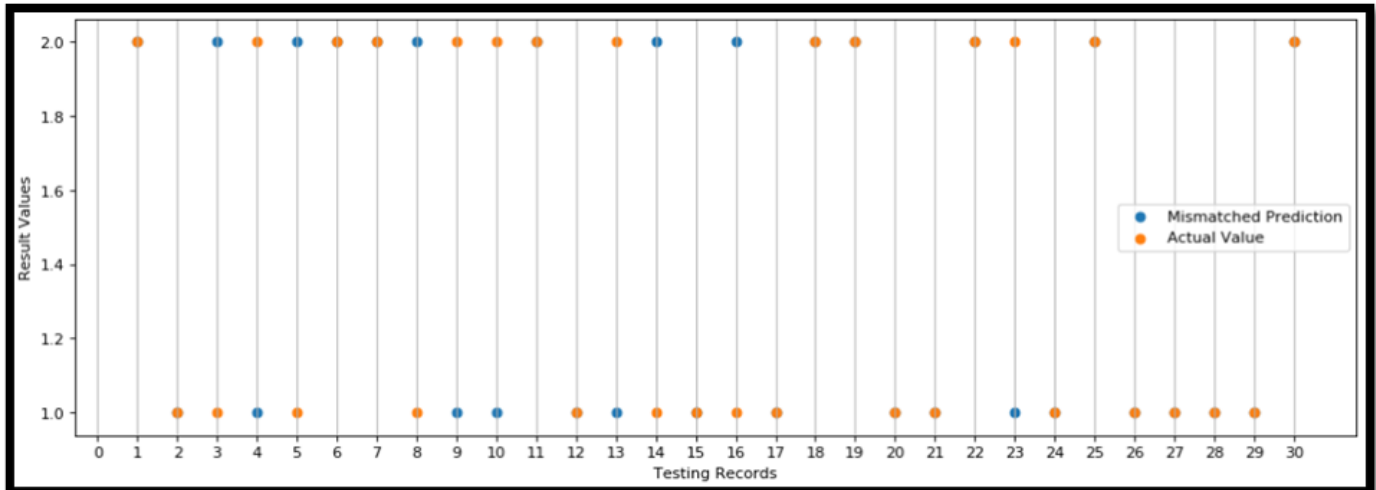


Chart -3: Scatter Plot for Heart Prediction using KNN

Overall, it is seen that out of the three algorithms used, Random Forest topped the accuracy with a result of 86.67%. This stands in support with findings of the authors in Paper [6] and Paper [7]. Hence the same is used to predict whether user has heart problem or not.

3.2.2 Accuracy of the Results for the Generalized Disease Prediction

In this case we have used 52 testing records those have been predicted using three different algorithms. The graphs and accuracy for these algorithms are shown in **Chart 4**, **Chart 5** and **Chart 6**.

A. Random Forest:

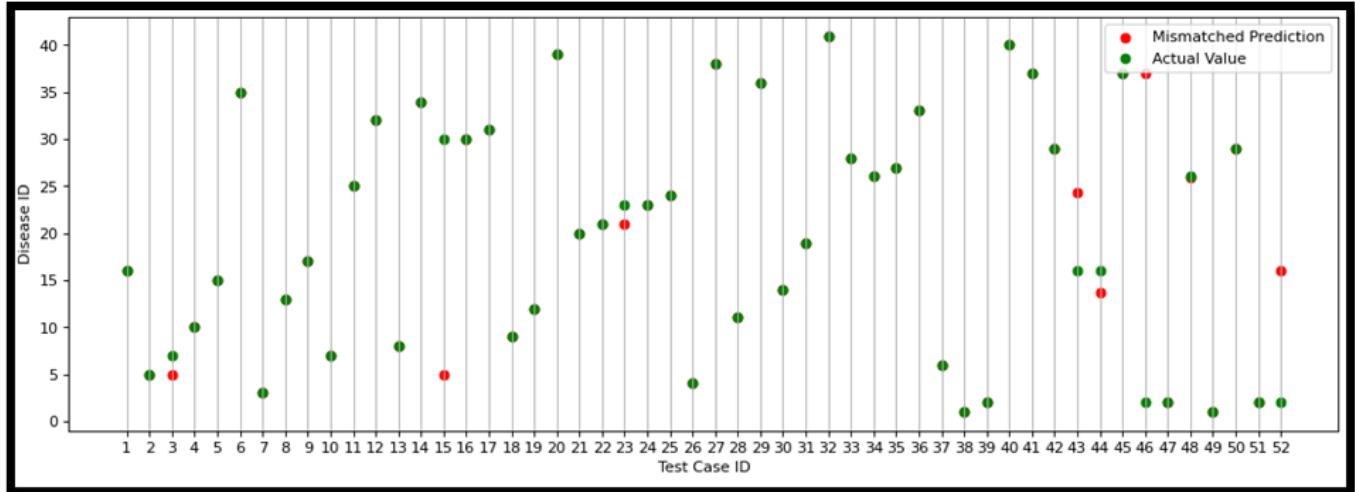


Chart -4: Scatter Plot for Disease Prediction using Random Forest

As shown in **Chart 4** it can be inferred that out of the 52 records tested, there are 7 incorrect predictions. So correct predictions are 45 and incorrect predictions are 7. Paper [5] has used Random Forest for disease prediction.

Therefore,

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Number of Records}} = \frac{45}{52} = 86.53\%$$

B. Naïve Bayes:

As shown in **Chart 5** it can be inferred that out of the 52 records tested, there are 4 incorrect predictions.

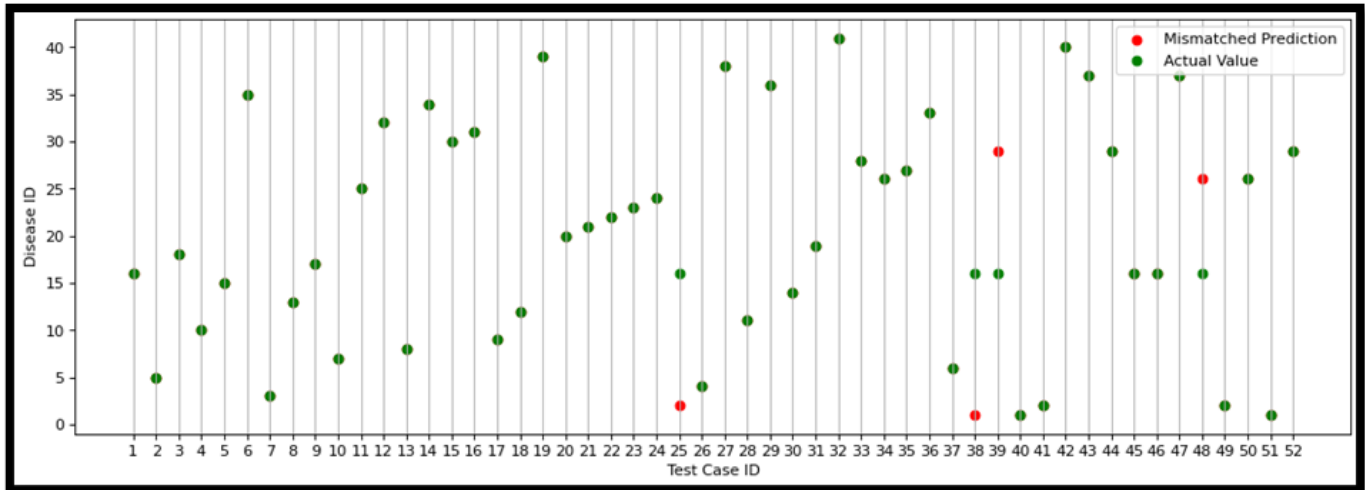


Chart -5: Scatter Plot for Disease Prediction using Naive Bayes

So correct predictions are 48 and incorrect predictions are 4. Use of Naïve Bayes can also be seen in Paper [4], Paper [7] and Paper [8] for disease prediction. Therefore,
 Accuracy = Correct Predictions/Total Number of Record
 = 48/52 = **92.31%**

C. K-Nearest Neighbor:

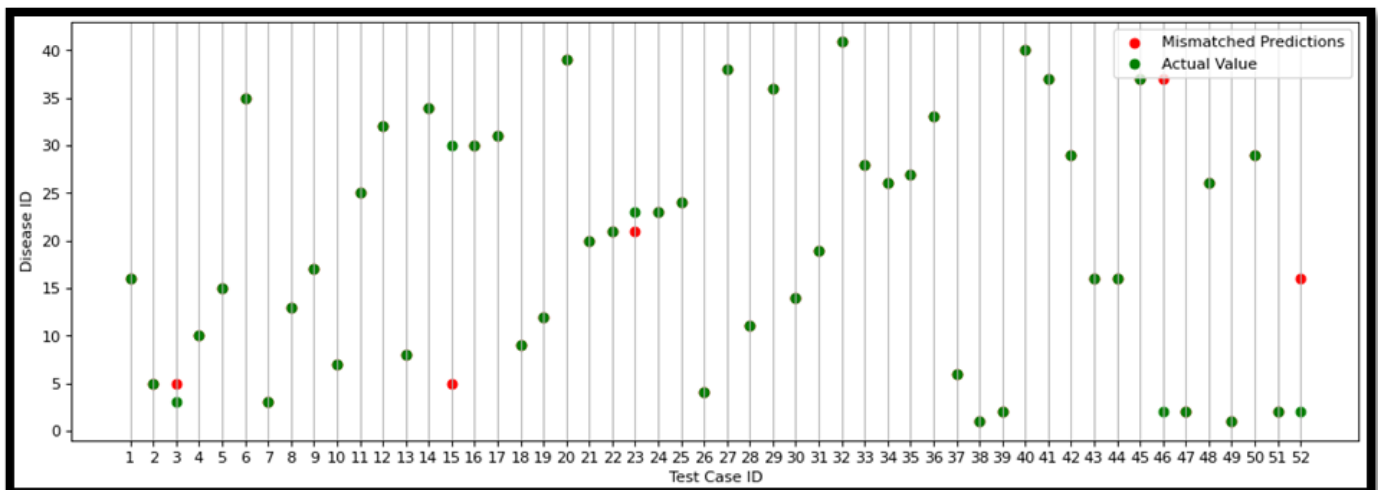


Chart -6: Scatter Plot for Disease Prediction using KNN

From **Chart 6** it can be inferred that out of the 52 records tested there are 5 incorrect predictions. So correct predictions are 47 and incorrect predictions are 5. KNN has also been used in research Paper [6], Paper [7] and Paper [9] and had almost similar results.
 Therefore,
 Accuracy = Correct Predictions/Total Number of Records
 = 47/52 = **90.40%**

Therefore, by looking at above results it was concluded that Naïve Bayes is best suited for this task as it topped the list with an accuracy of 92.31%. In Paper [4] the authors discussed how powerful Naïve Bayes algorithm can be and here we can see it performing best out of all other algorithms.

3.2.3 User Inputs and Results

As shown in Paper [2], the best way to implement such tools is by developing an interactive interface where users can enter their symptoms and get results. To make the system more personalized, user first needs to create an account in order to check the symptoms and diseases. After the account has been created the user can log in using his credentials. After the user has successfully logged in, there will be three options are presented before him:

- (i). Use heart disease predictor, shown in Fig 6.
- (ii). Use generalized disease predictor, shown in Fig 7 and Fig 8.
- (iii). Check check-up history, shown in the Fig 9 and Chart 7.

A. For Specialized Heart Check-up

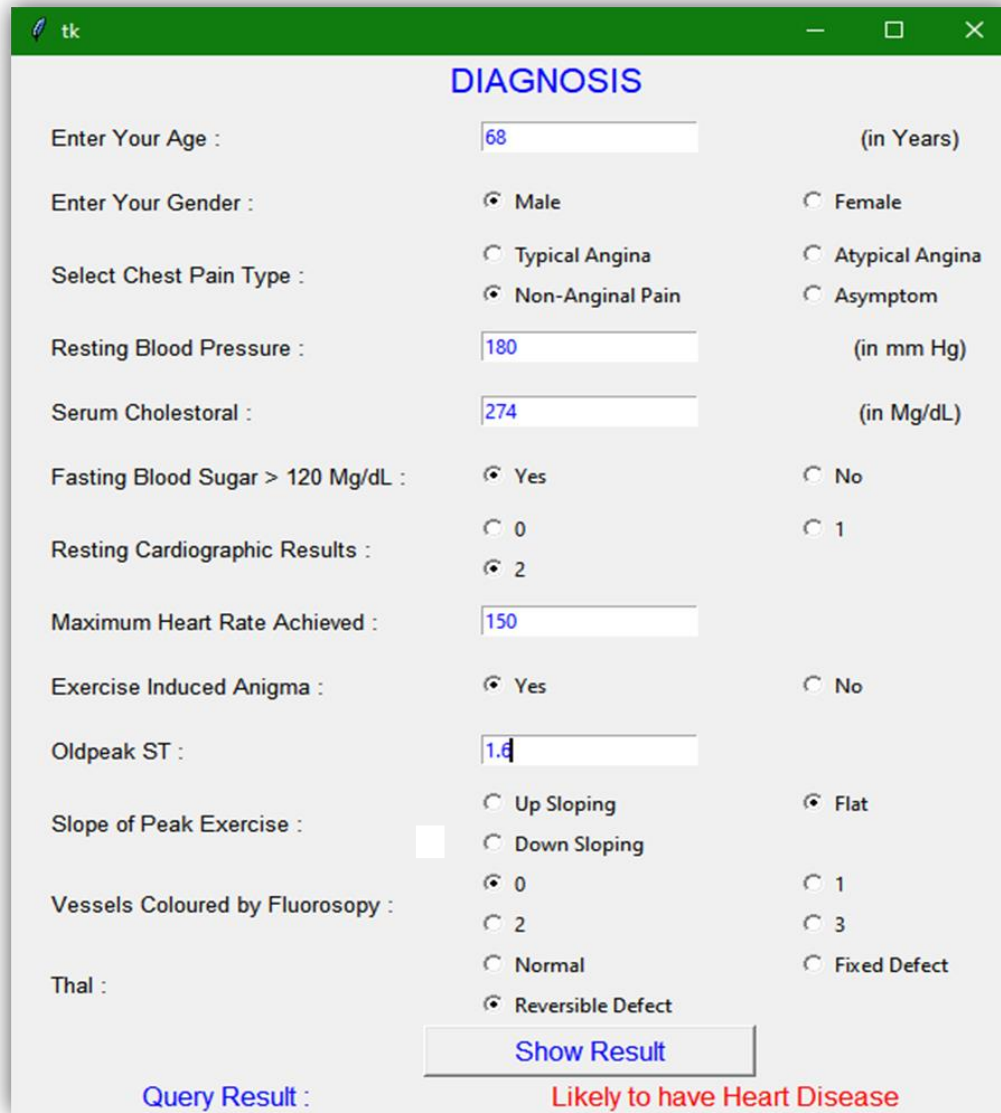


Fig -6: Interface for Input of Heart Parameters

As shown in Fig 6, the user will be required to enter parameters related to heart condition. These are essential in order to judge accurately. After entering the parameters, the user can see whether he is likely to be suffering from heart disease. The prediction is done using Random Forest Algorithm as it had highest accuracy when tested. Such results were also displayed using same algorithm in Paper [3]. If user is found likely to be suffering heart disease, it is advised to visit nearby hospital for other tests and treatment, if needed. More parameters can be added to improve the accountability of this prediction.

B. For generalized disease prediction

This is another option available for a generalized checkup that contains set of 42 diseases that are more common than others. This can help to judge whether the symptoms the person is suffering from are severe or not. Based on this the user can either visit doctor for treatment, or it is advised to take appropriate food and exercises, if severity is found low.

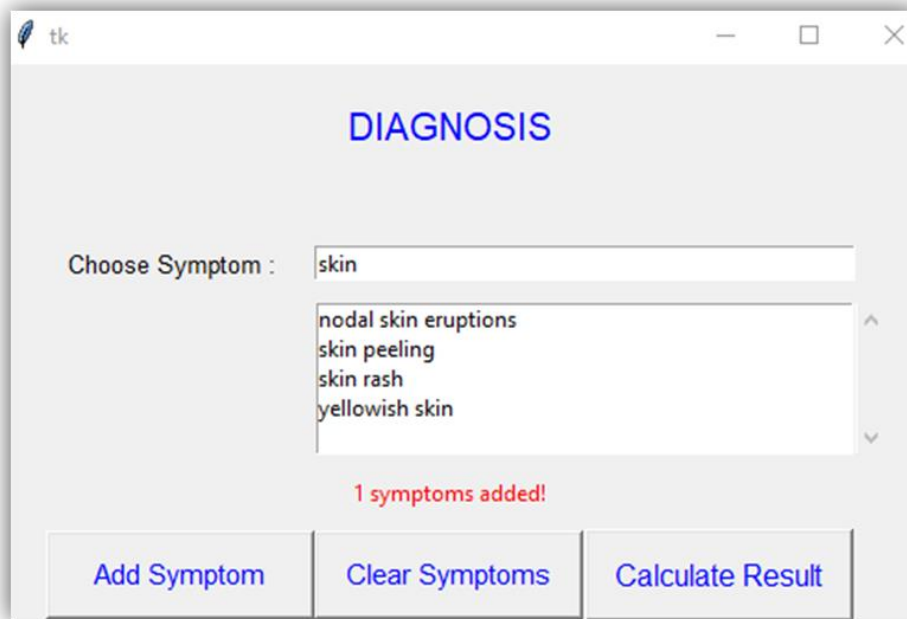


Fig -7: Interface for Symptoms Input

Fig 7 shows the interface used where the user can add symptoms. User only needs to type any keyword for the symptom and the suggestion will be popped in the list box. This implements search even for unformatted data as discussed in Paper [9].

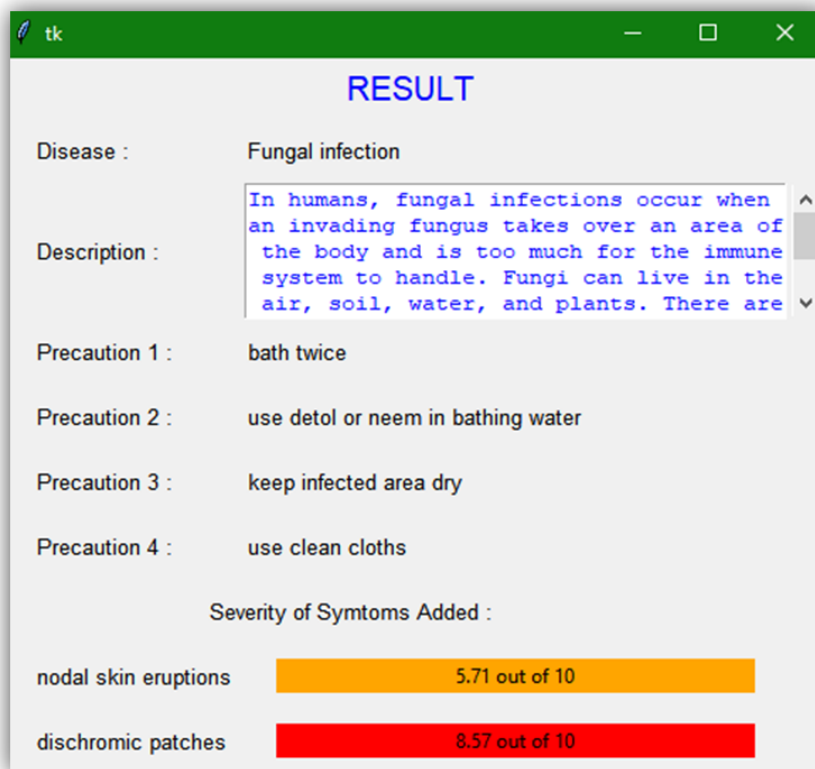


Fig -8: Results for User Disease Based on Symtoms

The user will be notified on how many symtoms have been added. There is also an option to clear symtoms if one has added it wrongly.

After the symtoms have been added by the user and the user has clicked the 'Calculate' Button, next result window is generated for which the system will use the symtoms given by user and predict what the cause might be. This is done using

previous data and there is a trained model in backend. The algorithm used here is Naïve Bayes. The advantages of Naïve Bayes have also been discussed in Paper [6]. Fig 8 shows the result of user inputs.

C. User History Feature and Insights

This section will be used to check previous check-up details. Insights will also be generated in background to help user understand when one's condition is worst and when the user is improving. This can be personalized even more as the data from previous check-ups can be used to predict what the main cause of health issues for the user is.

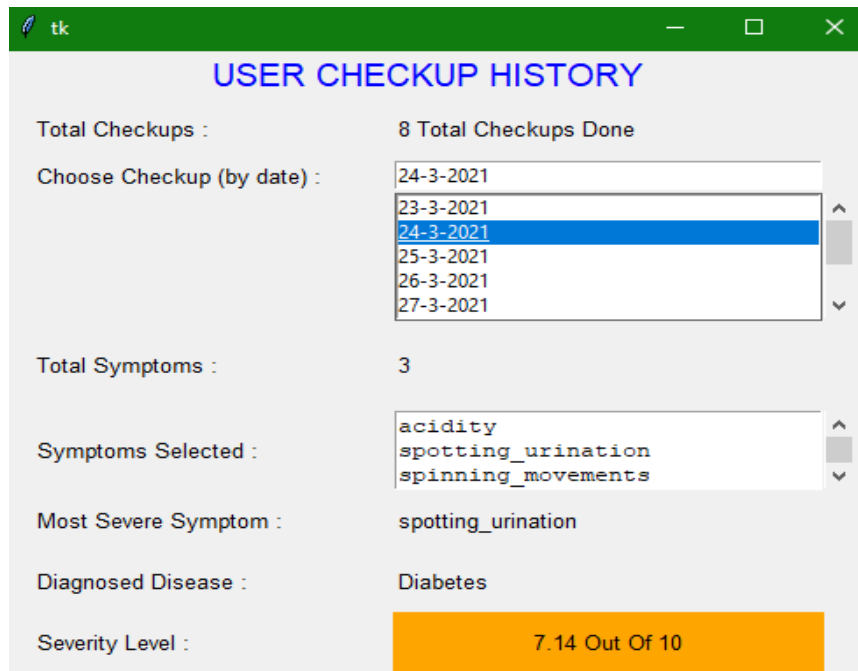


Fig -9: Interface for User History

In Fig 9 the basic idea will be to distinguish check-ups on basis of data. As discussed in Paper [1], personalization can significantly improve how effective a tool is as it can give more options to the user to make the experience according to what the user wants. Not just this, but also the how well described your information is can make huge difference. In the interface shown in Fig 9 only some important features like severity indication of symptoms, the symptoms themselves and overall severity is shown. This date-wise data can also be used to see which days the user was suffering from more severe symptoms, and when he was safe. Furthermore, as shown in Paper [6], more information can be later added as more research is done in this regard.

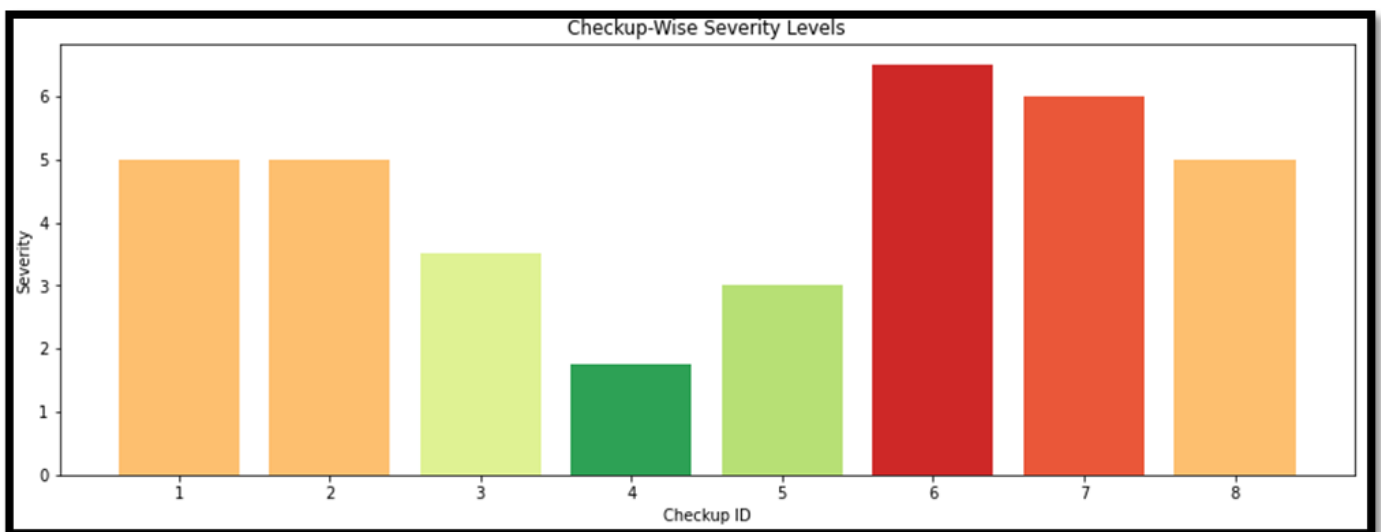


Chart -7: Plot for Severity of User Check-up History

Shown in **Chart 7** are the details such as when the user was improving. For example, in this case the condition was improving from first check-up till fourth checkup. Afterwards the condition became worse till sixth check-up. Hence the user habits can be improved by following plans or medication that the user was using from first check-up to fourth check-up. It can also be visualized that during the fourth check-up, the user was most healthy as severity of symptoms was lowest.

4. FUTURE WORK

The theme is ever evolving in nature. The degree of personalization a person can get in this platform is limitless. The symptoms that a user has encountered in past can be considered to predict the degree of care the individual needs. The number of insights generated by the system is directly proportional to the amount of data fed to the system. Hence integrating more data in terms of count of diseases and symptoms as well as the knowledge of what needs to be done in order to take care of the underlying disease can be increased as more data is integrated more insights can be generated using the check-up history of individual. The more personalized experience a user will get in both generalized and specialized predictions, the better results would be generated. One can also take this research work to the level where algorithm can collect data on its own in real time.

5. CONCLUSION

The main goal of this project was to facilitate the disease prediction in short time so that people won't ignore the symptoms just because they were being lazy. As the advancements of technology in this field is taking place at a much greater rate than ever before, the wider use of such prediction needs to be implemented. It is an essential thing required for everyone to be using as we are getting more and more careless about our health. The generalized approach is good at giving knowledge on what the underlying disease may be in the case of someone not much knowledgeable in this field, but the specialized sections are essential to those who have some knowledge about body parts, that they need to be worried about. Using both approaches it is essential to be as diverse as possible along with being more accurate in giving the desired information to the user.

6. REFERENCES

- [1] Khurana Sarthak, Jain Atishay, Kataria Shikhar, Bhasin Kunal and Arora Sunny, "Disease Prediction System" in International Research Journal of Engineering and Technology (IRJET), May 2019.
- [2] Tiwari Divyansh, Kumar Arpit and Tripathi Ayush, "Virtual Doctor" in International Journal of Advanced Research, Ideas and Innovation in Technology, 2019.
- [3] Rathi Megha and Pareek Vikas, "An integrated hybrid data mining approach for healthcare", in IRACST -International Journal of Computer Science and Information Technology Security (IJCSITS), Nov-Dec 2016
- [4] Sako D. J. S. and Palimote J., "A Medical Document Classification System for Heart Disease Diagnosis Using Naïve Bayesian Classifier" in International Journal of Applied Science and Mathematical Theory, 2018.
- [5] Patel Jaymin, Prof. Upadhyay Tejal and Dr. Patel Samir, "Heart Disease Prediction Using Machine learning and Data Mining Technique" in IJCSC, March 2016.
- [6] Adnan Muhamad Hariz Muhamad, Husain Wahidah and Rashid Nur'Aini Abdul, "Data Mining for Medical Systems" in Research Gate, May 2016.
- [7] Ramalingam VV, "Heart disease prediction using machine learning techniques" in International Journal of Engineering & Technology, March 2018.
- [8] Pingale Kedar, Surwase Sushant, Kulkarni Vaibhav, Sarage Saurabh and Prof. Karve Abhijeet, "Disease Prediction using Machine Learning" in IRJET, Dec 2019.
- [9] DK Harini and M Natesh, "Prediction of Probability of Disease Based on Symptoms Using Machine Learning Algorithm" in IRJET, May 2018.
- [10] R Sneha, S Monisha, C Jahnvi and S Nandini, "Disease Prediction Based on Symptoms Using Classification Algorithm" in Journal of Xi'an University of Architecture & Technology, 2020.