Project Report

on

Disease Diagnosis

Submitted as fulfilment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science And Engineering

By

NITIN SINGH-1809710065

NITIN KUMAR YADAV-1809710064

Under the Supervision of
Mr. T. Akilan
Mr. Mayank Dixit

Galgotias College of Engineering & Technology
Greater Noida, Uttar Pradesh
India-201306
Affiliated to

Dr. A.P.J. Abdul Kalam Technical University

Lucknow, Uttar Pradesh,

India-226031

Content

INTRODUCTION

Millions of people are treated in emergency rooms in hospitals every year. However, a significant proportion of the patients are non-emergency cases, forcing the hospitals to allocate medical personnel where it is not necessarily needed, creating a suboptimal use of staff and treatment of real patient emergencies. Moreover, there is a growing shortcoming of physicians in rural areas, leading to underserved patients, in particular, due to the demographic change and a growing number of elderly people .In the situation of COVID epidemic we all experience excessive lack of Doctors and due to which people with the minor health problem suffered alot for appointment and they also gone through the various Covid Test procedures for getting a basic disease prescription from Doctors, that motivates me to work on that problem and get its solution and one solution to overcome these problems in the future might be the use of artificial intelligence (AI) in healthcare as a driver for systems medicine and it also act as a filteration layer between doctors and patient . AI as speech recognition systems, such as Apple's Siri or Amazon's Alexa, have already entered our daily lives. They can be used as personal assistants or as an interface to interact and control smart homes. Moreover, speech recognition systems are also employed to support physicians in their daily routine in hospitals, e.g., to provide a computer-assisted documentation, i.e., they use speech recognition to translate speech to text and dictating patients' anamnesis results to fill in electronic health records. However, so far, speech recognition has not found its way into clinical decision-support systems. Today, clinical decision-support systems that are frequently used in clinical practice are rare, as they require expert knowledge and most of them have not been comprehensively clinically evaluated. However, there are some examples, including image analysis software for, e.g., magnetic resonance or computed tomography images, or expert systems that are used to predict treatment

options, e.g., in infectious diseases, which are fully integrated into daily clinical use.

We are trying to provide a user an interface through which they can connect to a AI model which is a mimic of a Doctor brain which can diagnose the disease of the patient using various AI tools after analyse the data provided by the patient that can be symtoms ,test reports, past disease records and many more. It is a like an interaction between a real user and a Machine as a doctor which give basic prescription and precaution and preffer medicines on the basis of predicted disease using that model. This systems are used by a patient or user in order to analyze patient data and to make predictions on possible outcome of a treatment or to provide additional data for a diagnosis or prognosis. This model are typically based on statistical and machine learning approaches, however, they are not coupled with speech recognition or synthesis to interact with a patient. The aims of the current model were:
1) The development of an AI able to interact with a patient (virtual doctor)
2) To demonstrate its usefulness on the automated prediction of disease on the basis of patient data.
3) Providing medicinal prescription and precaution and health tips.

Literature Survey

Disease prediction is most important for medical system to make the best possible medical care decisions.Incorrect decisions are likely to purpose suspensions in medical treatment or even loss of life.A number of disease prediction models are used in medical diagnosis system which are using data mining and machine leaning techniques like Bayesian classification, Decision Tree, Regression model, Neural

Network, Single best model,Ensemble model etc. In normal medical diagnosis system, it predicates the disease based on the patients symptoms and laboratory data before analyzing the disease. This prediction techniques give the good performance and with less accuracy using medical dataset. Neural Network is improving the observation capacity of information systems over the training of a limited number of neural network's nodes and collecting their results. Proposed system is enhancing the performance and training of neural networks for the classification with using cross-validation tool for the optimizing the network parameters and architecture. Using the artificial neural network (ANN) technique for disease prediction are compared below. Limitations of this scenario are as mentioned below:

I. System is only using one data set for validation which does not predictable enough to generate outcomes.

II. System is only exploring the common predictable performance of their models without considering the F-score and precision as measures.

III. Most studies do not provide statistical test results to demonstrate the level of significance of their experimental results.

IV. Most studies related to ensemble classifier do not compare the performance difference between individual classifiers and an ensemble classifier consisted of individual classifiers.

Already some researched in automated imaginative systems for medical applications is an essential and impressive. The classification of automated decision support system is a feasible to collaborating physician rapidly and accurate diagnose patients.Automated systems are useful in giving fast and accurate results. It is also helpful in reducing cost and time. It uses patient database to give enhanced results. The Fuzzy Min-Max network for medical diagnosis system shows

how the Classification, Regression Tree and Random Forest models are integrated to make a hybrid intelligent system. Rule Extracting is the pros of the CART and it is in a tree based structure. It is not more flexible to perform same accuracy on medical data samples. Absences of capability of predictions is the pros of the FMM. In medical system, accuracy of the DSS is more decisive. The prospective method is not gain the high precision, sensitivity, and specificity rates, but still to contribute description for its prognosis in the structure of a decision tree. Using Fuzzy Hierarchical Approach to Medical Diagnosis we can improve the results by following ways:

• Complexity in diagnosis process.

• Simple fuzzy logic which does not provide hierarchical structure.

• Uncertainty occurred by different diagnosis system.

• Problem occurred in other diagnosis system where grammatical labels comprehend to actual code in a period a numbers of values sensible process ca approach.

Using Limitation of the profession and aspect for medical diagnosis system using machine learning we can improve the result as it:

• Provides an analysis of the automated data scrutiny.

• Significance the naive Bayesian,neural network,Decision Tree.

• Specific requirement for machine learning systems.

- Good Performance
- Dealing with missing data.
- Dealing with noisy data.
- Reduction of the number of tests.

Exciting data mining and Machine learning techniques which are used in medical field.

1.Support Vector Machine:

SVM method works based on particular disease and gives the accurate result also. When it was used for another disease, it was not produced the accurate result because for modification in current algorithm is required for particular disease.

2.Naïve Bayes

• Probability based checked and found the disease.

• Using the class conditional probability.

• Which disease of probability high than it bias on that disease and give the result of the highly probable disease.

3.Decision Tree

• Complexity increases.

• Time consuming process.

4.Random Forest

• Forest is made using more than one decision tree

• Splitting criteria of decision tree is random attribute selection.

• Random vectors are sampled independently and with equal distribution among all the trees.

• The class selection process includes voting of every tree and majority voting class is returned.

5.Clustering Method

• Simultaneously change cluster based on symptoms.

• It's not give the accurate result for the number of diseases.

6.Logistic Regression

• Recursive Procedure.

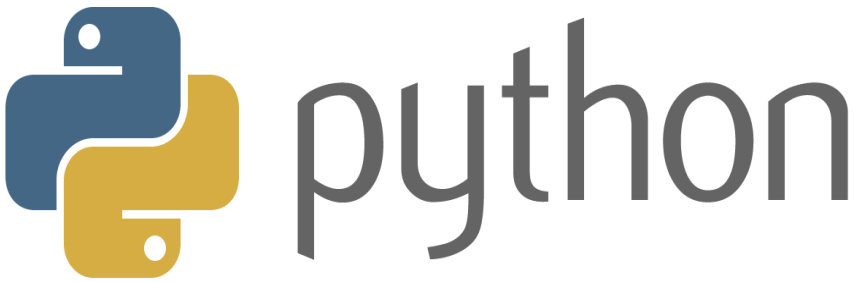
•Time consuming process.

7.Backpropagation

• Predefined hidden units.

• Time complexity increase.

Language platform



Python was created in the early 1990s by Guido van Rossum at Stichting
Mathematisch Centrum (CWI, see http://www.cwi.nl) in the Netherlands
as a successor of a language called ABC. Guido remains Python's
principal author, although it includes many contributions from others.

In 1995, Guido continued his work on Python at the Corporation for
National Research Initiatives (CNRI, see

http://www.cnri.reston.va.us)
in Reston, Virginia where he released several versions of the software.

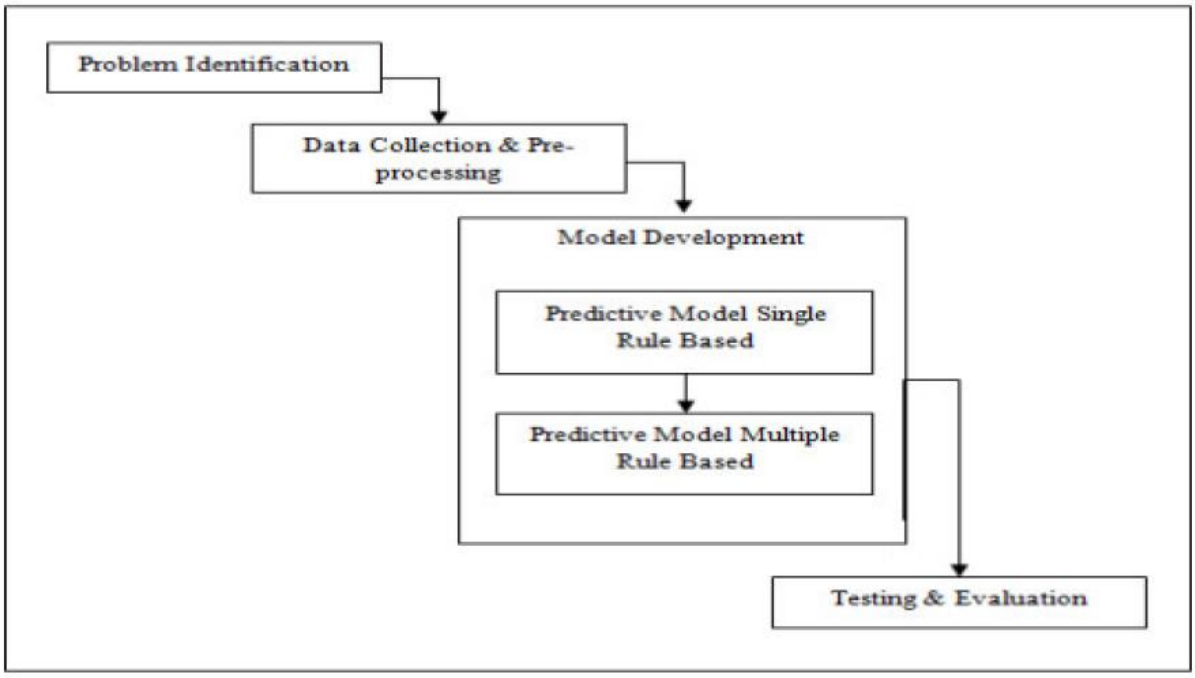
In May 2000, Guido and the Python core development team moved to
BeOpen.com to form the BeOpen PythonLabs team. In October of the same
year, the PythonLabs team moved to Digital Creations, which became
Zope Corporation. In 2001, the Python Software Foundation (PSF, see
https://www.python.org/psf/) was formed, a non-profit organization
created specifically to own Python-related Intellectual Property.
Zope Corporation was a sponsoring member of the PSF.

All Python releases are Open Source (see
http://www.opensource.org for
the Open Source Definition). Historically, most, but not all, Python
releases have also been GPL-compatible; the table below summarizes
the various releases.

Proposed Methology

The proposed model works as a client server based system which is implemented on a websites named as Docmac which is a AI doctor.

We are trying to provide a interface to the user in frontend by making a website in which we implement our disease diagnosis model which can take input through the website by GET/PUT request method to our model which can predict the result and then an output page is showed to the user which contain all the details about the predicted Disease and contain the respective basic prescription and precaution.



Tools Used:

1.HTML
2.CSS
3.JAVASCRIPT

4.PYTHON
5.FLASK, SKLEARN, PANDAS, Numpy,Json
6.BOOTSTRAP

For frontend website we use HTML ,CSS and JAVASCRIPT for designing and making our website more dynamic and easy to handle for users.
The model is implemented at the Disease Detector section of the website which contain input forms where user can input symptoms that they feels in their body. These symptoms get submited and get by model at backend by request method.

We are using Flask module of python as a web application framework for handling our website and web page routes, it is a great tool for web development using python.

Our machine learning model code is written in python used as a programming language and also using various python module such as Pandas,Numpy,Sklearn for data analysis and model creation and Matplotlib for visualisation of data.

```
1]: import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import LabelEncoder
    from sklearn.linear_model import LogisticRegression
```

Dataset is very crucial and important part in model creation because all the prediction are based on the quality of data that we collected from various sources.

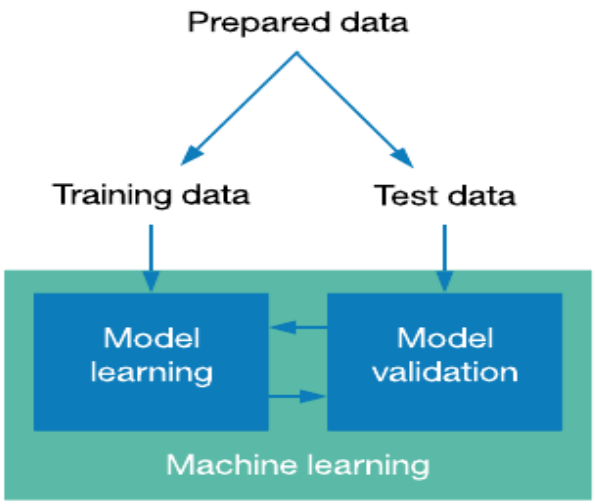
```
[4]: df=pd.read_csv('dataset.csv')
     df
```

t[4]:

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	Na
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	Na
2	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	Na
3	Fungal infection	itching	skin_rash	dischromic _patches	NaN	NaN	NaN	NaN	Na
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN	NaN	NaN	NaN	Na
...	...	...	...	...	...	...	...	...	..
4915	(vertigo) Paroymsal Positional Vertigo	vomiting	headache	nausea	spinning_movements	loss_of_balance	unsteadiness	NaN	Na
4916	Acne	skin_rash	pus_filled_pimples	blackheads	scurring	NaN	NaN	NaN	Na
4917	Urinary tract infection	burning_micturition	bladder_discomfort	foul_smell_of urine	continuous_feel_of_urine	NaN	NaN	NaN	Na
4918	Psoriasis	skin_rash	joint_pain	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	NaN	Na
4919	Impetigo	skin_rash	high_fever	blister	red_sore_around_nose	yellow_crust_ooze	NaN	NaN	Na

4920 rows × 18 columns

The dataset is furthur divided into two parts one is for testing and another one is for training the model and these parts are named as testing and training datasets.




```
X_train, X_test, y_train, y_test = train_test_split( X, y,
test_size=0.33, random_state=42)
```

Training dataset is used for training the machine learning model on the basis of which the model create a mapping function which maps the training values with output. Testing dataset are used for testing the prediction acurracy of the model .

We apply various machine learning algorithm for the classification such as Support Vector Machine, Logistic Regression and Decision Tree Classifier and after training and testing our model using these algorithm we find score for each algorithm and on this basis we find which algorithm is best for detection of disease.

Score and cross result using DecisionTree Classifier

```
In [5]: clf1  = DecisionTreeClassifier()
        clf = clf1.fit(x_train,y_train)
        print("score:",clf.score(x_train,y_train))
        print ("cross result========")
        scores = cross_val_score(clf, x_test, y_test, cv=3)
        print (scores)
        print ("mean:",scores.mean())
        y_pre=clf.predict(x_test)
        y_pre0=clf.predict([14])
```

```
score: 1.0
cross result========
[0.97232472 0.97227357 0.97597043]
mean: 0.9735229052845057
```

Score and Cross result using Support Vector Machine

```
In [7]: model=SVC()
        model.fit(x_train,y_train)
        print("for svm: ")
        print("score:",model.score(x_test,y_test))
        print ("cross result========")
        sx = cross_val_score(model, x_test, y_test, cv=3)
        print(sx)
        print("mean:",sx.mean())
        y_pre1=model.predict(x_test)
        y_pre10=model.predict([14])
        #print(y_pre10)
        #y_test
```

```
for svm:
score: 1.0
cross result========
[1. 1. 1.]
mean: 1.0
```

Score and Cross result using Logistic Regression

In [8]:
```
detector=LogisticRegression()
detector.fit(x_train,y_train)
print("score:",detector.score(x_test, y_test))
print ("cross result========")
sco=cross_val_score(detector,x_test,y_test,cv=3)
print(sco)
print("mean:",sco.mean())
y_pre2=detector.predict(x_test)
y_pre20=detector.predict([l4])
#print(y_pre2)
#y_test
#np.array(range(132))
```

```
score: 1.0
cross result========
[1. 1. 1.]
mean: 1.0
```

We are using a classification technique for disease prediction for that we use Logistic classification algorithm of ML which can classify all the diseases on the basis of symptoms. It is a supervised learning algorithm which need data for analysis and reaches to the best prediction.

CONCLUSION

In this project, we have proposed methods for diagnosing disease in patients using machine learning techniques. The three machine learning techniques that were used include SVM, Logistic Regression, decision tree classifier. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. Logistic Regression was the model that resulted in the highest accuracy . Comparing this work with the previous research works, it was discovered that ANN proved highly efficient.

Future scope and enhancement

The important and crucial part in this project is the dataset because accuracy of the prediction are totally based on these data. So in future we can improvise dataset for better predictions and there are various models too which can be implemented on the data for making model more accurate.
In future we can do lot of work in this project for taking it to advance level :

Using Medical test reports for analysis

We can use test reports of the patient as one of feature for getting more insights of disease on the basis of data collected through patient reports which can make our model more effective and accurate. For that we can use Natural language processing for analyzing report and extracting useful data from it.

Implementation of Chatbot system in website

We can also implement a Chatbot like system in our website, respective of taking input from the patients as a form we can use text to speech system like a Chatbot which can interact to patient and trying to make it look like a realistic patient Doctor interaction.

REFERENCES

1. https://www.kaggle.com/
2. https://getbootstrap.com/
3. https://scikit-learn.org/stable/
4. https://numpy.org/
5. https://pandas.pydata.org/
6. https://archive.ics.uci.edu/ml/index.php