IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Voice Interactive Bilingual Smart Healthcare Chatbot Using NLP

[1]Ankit Kumar Singh
Department Of Computer Science and Engineering
SGSITS, Indore

[2]Mrs. Neha Mehra
Department Of Computer Science and Engineering
SGSITS, Indore

Abstract— Medical and healthcare-related material on the Internet has expanded dramatically in recent years. On the one hand, recent research found that the number of internet users searching for health-related information online is increasing. Artificial intelligence (AI) technology IS now extensively employed to assist with knowledge acquisition and decision-making in a range of sectors. Particularly in terms of health information systems, AI has a lot to offer. The healthcare sector has recently seen a growth in the importance of research and production related to symptom-based sickness prediction. Several scientists as well as organizations have demonstrated an interest in analyzing and developing innovative techniques for rapidly and correctly predicting diseases using modern computer technologies. We provide a model for evaluating the effectiveness of merging In this research, we integrate machine learning (ML) & natural language processing (NLP) techniques into an illness prediction system. Pattern recognition in medical data is one application of machine learning. As a result, it is capable of accurately forecasting sickness. We employed Machine Learning methods, Python Programming with Jupyter Interface, as well as a dataset obtained from Kaggle to accomplish Disease Prediction based on Symptoms. The prediction power and illness categorization patterns of the NLP and ML models differed. We examined feature correlation using a confusion matrix. The proposed modes are 99% accurate. Our novel ML models obtained great illness prediction efficiency via disease categorization. This research will be valuable in illness prediction and diagnosis.

.

I. INTRODUCTION

The term "smart healthcare system" refers to a healthcare system that uses technological advancements such as artificial intelligence (AI), blockchain, big data, cloud/edge computing, and the internet of things (IoT) to create numerous intelligent machines that connect healthcare respondents and enhance healthcare quality [1]. When discussing smart healthcare, the three primary categories of participating that may be divided into separate groups are now the general public, healthcare service providers, & third-party healthcare involvement. Smart homes, smart hospitals, intelligence research and enhancement for life science, patient care, population health, rehabilitation treatment, and a variety of scenarios are some pertinent instances of smart healthcare scenarios in relation to the participant. Natural language processing is identifying essential features in a common language and extracting meaning from unstructured or spoken written input using computational models. The field of natural language processing (NLP) calls on expertise in artificial intelligence, computational linguistics, and many other areas of machine learning. Figure 1 depicts the primary players in the smart healthcare industry, along with example situations and upcoming technology.

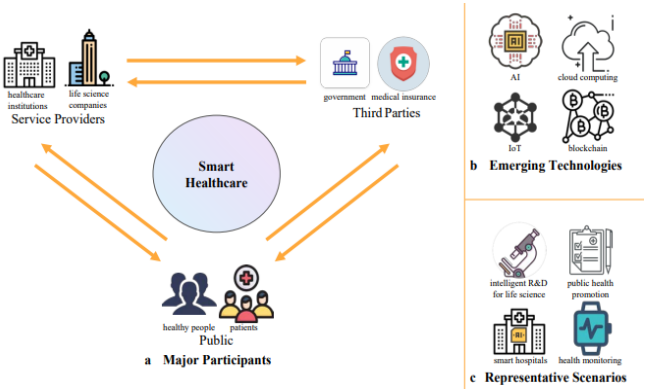

Figure 1: Smart Healthcare technologies [2]

Natural language processing (NLP), which is connected to human-machine & human-human communication, is used in smart healthcare to process text data. The text data can be separated into clinical text as well as other text data. The majority of the clinical text comes from electronic health record (EHR) systems and emanates from all clinical settings. These records contain a variety of data, such as radiological reports, electronic prescriptions, and medical notes. The term "other text data" refers to any text that is found in other types of healthcare contexts, such as questionnaires used for population screening and publications used for evidence-based referencing. All aspects of health care involve some form of communication, whether it be between patients and healthcare professionals during clinical investigations or between patients and drug treatment robots during neurological rehabilitation; this communication is often preceded by applications like machine translations as well as interface design for restoration robots[2][4][5].

A growing variety of medical as well as healthcare-related things are now accessible via mobile and web apps. Several machine learning research initiatives start by using information acquired from online platforms like social media, forum chats, and a range of other resources to build AI-supported medical recommendation applications. The outcomes of the research were quite promising, and these AI applications may provide valuable recommendations or even pre-diagnose advice based on very basic information, including such disease as well as symptom relational databases. Furthermore, the bulk of the work lacks a simple framework for preserving data and utilizing ML algorithms on reliable data. The use of machine learning,as well as clustering for cancer prediction, a naive Bayes classification for the prediction of dermatology illnesses, anda naive Bayes classifier for the prediction of the presence of swine flu, have all been researched in the past, and all look to give very reasonable results. The bulk of this research, however, concentrates on one or a few illnesses or medical issues that are crucial in the healthcare industry.Much research, however, presented unique ways of forecasting greater disease based on symptoms[6].

The empirical study of algorithms & statistical models known as machine learning (ML) enables computers to do tasks efficiently without the need for explicit instructions. As a portion of artificial intelligence, it is regarded as such. Machine learning techniques create a mathematical formula that can generate predictions or judgments without being explicitly instructed to do so by using sample data, also referred to as "training set." When constructing an algorithm with explicit instructions for achieving the goal is problematic, machine learning approaches are used in industries including email classification, detection of network intrusions as well as computer vision [6].

This analysis is organized as follows: Section 1 offerings an overview of the topic and background information. Section 2 portrayed the related works in disease prediction using machine learning and NLP areas. And the materials and methodology are designated with the evaluation criteria of different classifiers in Section 3. Moreover, the performance findings and discussion are illustrated in Section 4. Discussions of final thoughts and directions for further research make up Section 5.

II. Related Work

This section provides related work on medical health in NLP and machine learning, using various tools and technologies like deep learning and machine learning. Some related works are discussed below:

Recently, hospital care has become more available to more people because of technological advancements. Artificial intelligence and machine learning, including neural networks have all been of great use in the medical industry. People in today's fast-paced society often put off taking care of their health, which may be disastrous. The symptom-based illness prediction software is a viable solution to this issue. Hiba Hussain[7] focuses on instantaneous and precise illness prediction for consumers based on symptoms, as well as in-depth pathology report interpretation. A combination of NLP and ML algorithms is used to create the illness prediction chatbot. Researchers have employed the decision tree and indeed the K-Nearest Neighbor methods in illness prediction (k-nearest neighbors). These methods' results are compared to choose the most reliable model. Our findings show that KNN has a higher accuracy than Decision Tree at 95.74%.

In studies related to Doan *et al.* [8], employing methods from natural language processing (NLP), researchers tested a strategy for deriving causal links from tweets. The words "stress," "insomnia," and "headache" were the primary focuses of our attention in this section. To extract causal data, researchers came up with a collection of lexico-syntactic patterns based on dependency parsers' outputs. A massive dataset that included 24 million tweets was used for this study. According to the findings,

their methodology obtained an overall accuracy that ranged between 74.59% and 92.27% on average. Examining extracted relationships showed intriguing insights on topics linked to health on Twitter.

Subiksha [9] provides an application-oriented service-sharing model that generalizes knowledge processing and mapping to connect healthcare systems. Based on the model, a presentation is made of architecture for distributed deep learning that can easily integrate several healthcare databases and systems transparently. The necessary inputs are provided in phrases processed using NLP, which are then transformed into queries. This process is predicated on producing a supervised clustering that emphasizes semantically significant ideas and connections, which will then be connected utilizing causal links.

Dascalu *et al.* [10] describe the design and implementation of a smart platform for dosing drugs, including checking for probable interactions, establishing dosing notifications in a calendar, and adding medicines to a user's phone. Using NLP methods adapted for the Romanian language, the system compiles leaflets from Biofarm and HelpNet and makes them fully searchable using Elasticsearch. To aid in users' treatment and guarantee self-education, a user-friendly software device was built and made accessible on iOS and Android devices and internet browsers.

Yang *et al.* [11] introduce an NLP tool for standardizing Pack-Year unit measurements of smoking-related metrics extracted from clinical notes. An NLP systems with a two-tier rule-engine structure was developed, and 200 patient medical records from low-dose CT scans used for lung cancer screening were reviewed. The testing dataset showed the greatest results for natural language processing system, with F1 scores of 0.963% for a liberal assessment and 0.946% for a severe evaluation.

Researchers Desai, Gururaj,and Prakash [12] constructed a multi-label classification model to label plain text. The data for this model came from an Automatic Speech Recognition (ASR) model. The primary objective of the research is to build an accurate model that can make educated estimates about the tags associated with plain text and reach an accuracy rate of 82% using SVM.

Yu *et al.*[12], to extract DR-related themes from clinical narratives, researchers investigated two natural language processing methods considered to be state-of-the-art. These models are called BERT and Roberta. Four transformer-based NLP models were trained to extract clinical ideas after the researchers created annotation criteria, annotated a DR-corpus made up of 536 photo reports, and trained the models. The BERT model trained using the MIMIC III dataset got the best stringent and lenient F1-scores of 0.9503 and 0.9645, respectively, according to the results of the experiments.

In 2016, [13] prince and Thomas suggested a system for predicting the onset of cardiac disease by using data mining methods mostly centered on the KNN and ID3 algorithms. In this case, though, the dataset's provenance is unclear. Predictions were found to be accurate 80.6% of the time.

Technology's positive impact on healthcare is far-reaching. Machine learning algorithms have been useful for bothphysicians and patients by creating a real-world test population for the latter to use. AI's Natural Language Processing subfield helps medical professionals glean useful information from a patient's utterances. Predictions in machine learning are based on mathematical models that examine a variety of input data. As they run, ML algorithms improve and learn. These algorithms improve in precision when more information is added to them. Disease prediction employs a variety of supervised techniques.

III. RESEARCH METHODOLOGY

In recent years, the volume of data generated in the medical and healthcare sectors has increased dramatically. Big data analysis using ML is needed to manage this data and gain information from it to address a wide range of medical and clinical problems [14]. Numerous studies have demonstrated the significant improvement in performance that ML techniques have brought to medical categorization challenges. On the other hand, supervised learning-based approaches are among the most useful for scientific study and practical applications in clinical settings[15]. The primary purpose of these efforts is to enhance disease symptom prediction and diagnostic methods.

We will construct an effective framework that may be utilised to predict common diseases or medical issues based on symptoms that have been identified, looked into, and analyzed. This will be accomplished by contrasting the suggested framework with current solutions. The study provides forecasts for 42 medical disorders using natural language processing & machine-learning methods.

For the beginning of this work, First, we will discussthe data set part in the data set part we have read a CSV file,CSV file contains a list of diseases and their symptoms then we have converted those things into the form of the dictionary as per our ace of use than we have to covert those things into panda file than after pre-processing of data we have converted data string into a lower format. Rest information i.e. index or disease equivalent we have saved in the dictionary. Now in the training of the neural network part, we have used the autogluon library,autogluon library is a collection of many machine learning models in the first part weread the data set and saved it using the h5 file and loaded the dictionary using the pickle file.For classification, we will create an object of the autogluon model where df(x) is strange in our dataset whose column is our feature vector and rows contain 01 01 information and labels are stored in y we or we have started training our model network uses different models,models such as Fast AI, Light,GBM model, Random Forest, Decision tree, XGBoost, NeuralNetTorch,LightGBM Large, Catboost and many other models as mentioned in an above diagram than after training it creates their ensemble.In an ensemble what we will do is suppose we have to predict any disease based on its symptoms so now all those models which we have trainedwill pass it through all of them now after the result we will observe which disease is getting the most favor in all those models.Now in the testing

phase,we are predicting the top two diseases(For this we are using the softmax function) after loading the data set and dictionary now we are using the trained neural network where we are predicting the top two diseases based on the symptoms then we are giving explaining about the diseases and precautions that are needed to be taken. After predicting the disease, we are using practo.com so that users can get appointments withdoctors related to the diseases.

The structure is made up of five basic parts all working together:

A. Data Collection

A data set[1] is chosen from Kaggle which contains a list of diseases and symptoms total of 4921 rows that can classify 42 diseases.We have a filtered data set that contains symptoms and diseases corresponding to the symptoms. Some columns list the illnesses, their symptoms, the preventative measures that should be performed, as well as the weights of the diseases. Cleaning up this dataset is simple and may be done in any language that supports file management. To utilize this data effectively, the user needs just grasp the concept of rows and columns.

B. Data Preprocessing

Data preprocessing refers to any type of processing done on raw data to prepare it for another data processing method. Data preparation is one component of data preprocessing. It has historically been considered one of the most significant first steps in the process of data mining.In this research, we used a data preprocessing technique the Removed all the special characters,and white spaces from the data, Converted every word into a small alphabetand Creating a dictionary for disease to index and index to disease mapping.

C. Feature Selection

The feature is quite significant in terms of its impact on data processing. Feature selection is the process of narrowing down your model's inputs to only the most important information and excluding irrelevant or distracting details. It's when your machine learning model makes feature selections on its own depending on the issue you're attempting to address. High-correlation features are more linearly reliant on one another and as a result, provide a virtually identical impact on the dependent variable. Features having a strong correlation or substantial correlation with the dependent variable may be picked out with the use of this visual aid in the feature selection process.

D. Data Splitting

This method splits the whole dataset in half, creating two independent collections of information. After the data was cleaned and prepared, I split it in half, dividing the training set by 80% and the test set by 20%.

E. Classification With ML

This study will provide a mechanism for predicting illness symptoms using categorization methods. The ML classification method takes in training data and then utilizes it to create predictions about the likelihood that incoming data will fit into one of several preset classes. Several classification strategies were used for medical databases and diagnostic problems like illness symptom prediction. In this research, we have used 12 machine learning algorithms like LightGBM, and LightGBMXT. LightGBMLarge, NeuralNetTorch, XGBoost, CatBoost, WeightedEnseble_L2, ExtraTreeGini, NeuralNetFastAI, RandomForestEntr, ExtraTreeEntr,and RandmForestGini, some proposed techniques are described below:

1) LGBM

LightGBM [16] is a powerful implementation of GradientBoost that addresses the efficiency as well as scalability issues that plague GradientBoost when dealing with large datasets or high-dimensional feature spaces. All of the available data samples must be sifted through each feature inside GradientBoost, and all of the potential split points must be taken into account. The GradientBoost algorithm's splitting strategy utilised at each tree node is thus the most labor- and time-intensive component. In this method, gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) should both be used as sample techniques. Using the GOSS technique, only a small subset of data with significant gradients is considered when deciding where to draw the divide.

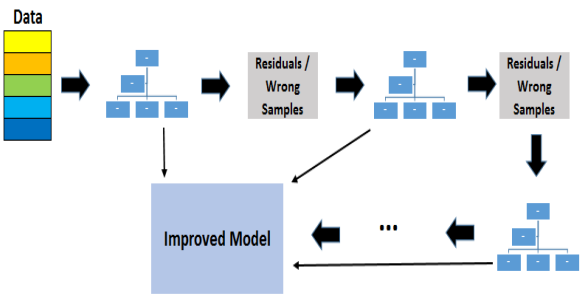

Figure 1:Understanding the LightGBM

[1] https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?resource=download

2) XGBoost

A model called XGBClassifier handles classification in XGBoost. A scalable, distributed, generative adversarial network (GBDT) machine learning system is known as "extreme gradient boosting," or simply "XGBoost." It enables parallel tree boosting and is the go-to machine learning tool for problems like regression, classification, & ranking. It is a highly scalable training strategy that prevents overfitting by using sequentially-built shallow decision trees to provide reliable predictions[17].

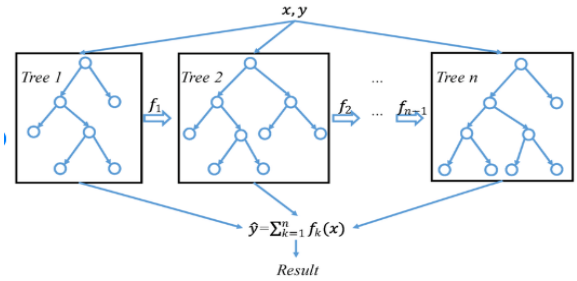

Figure 2: A general architecture of XGBoost

3) CatBoost

CatBoost is a brand-new gradient boosting technique for decisions trees that can train on categorical data. CatBoost is an AI learning system created by Google Brain. Specialists at Yandex develop and utilize it for a wide range of things: search, recommendation systems, administrative assistants, autonomous cars, weather forecasting, and more. The training process may be sped up via CatBoost, which leverages both GPU and CPU implementations. CatBoost divides tree formation into two stages, just like most other gradient-based boosting techniques. Binary and multi-class problems are only two examples of the sorts of problems that CatBoost may be used to solve [18][19].

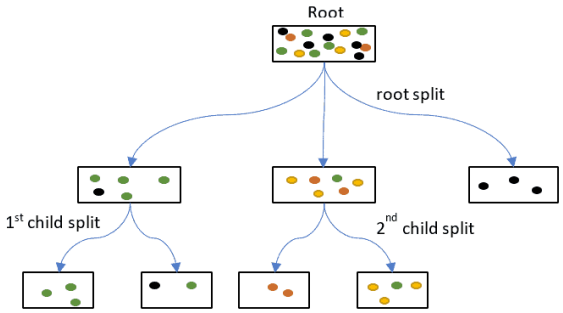

Figure 3: Example of CatBoost algorithm

4) Extra Tree

Decision trees are constructed using the traditional top-down method by the Extra-Tree classifier. These techniques substantially randomize the assignment of features and cut-point while partitioning tree nodes. If you push it too far, it will produce trees whose structures have nothing to do with the values in the training set. Only two key distinctions separate this method from similar tree-based communal processes: First, Instead of using a bootstrapped subset of the training data to construct the trees, it takes the whole dataset and randomly divides nodes within it. The average of all of the trees' projections is used to calculate the final prediction. It is anticipated that an extra-trees classifier may decrease variance more successfully than competing techniques since it uses complete randomization of the cut-point & attributes in addition to ensembles averaging. To reduce bias, all of the actual training data instead of bootstrap copies are used. There are several advantages to this method when it comes to the efficiency with which it processes computations(Ampomah, Qin, and Nyame, 2020).

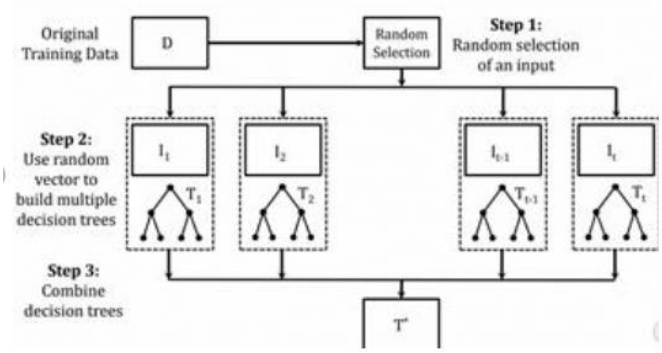

Figure 4:Visual Representation of Extra Trees Classifier

5) Random forest

RF is an enhanced version of the DT model that uses a supervised learning technique. Two main ideas contribute to the model's randomness. One issue is that during model training, certain trees may use the same data more than once because of the random way in which the data is distributed across the trees. The goal here is to reduce the model's variance and, by extension, the dispersion in the projected outcomes' scores. The second idea is to use a minimal selection of characteristics when partitioning the trees' nodes. This is done to avoid overfitting, which occurs when a model artificially inflates its predictions using the training data. For RF predictions, we employ bootstrap aggregation, which takes the average of the tree forecasts to get the overall data class. An advantage of RF over DT is that it uses several trees to generate predictions, each trained on a unique set of data and equipped with a unique set of characteristics.

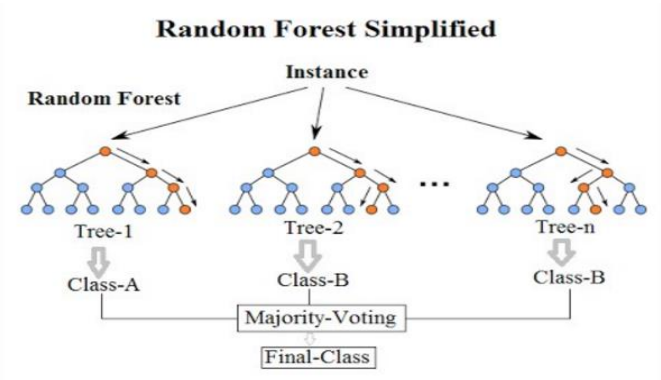

Figure 5: Random forest structure

6) Neural Network

The concept of neurons in biology is the inspiration for the artificial neural network (also known as an ANN). In the brain, cellular structures called neurons carry out several functions. Neuron theory is essential to understanding neural networks. There are essentially four primary components of a neuron. Dendrites, nucleus, soma, and axon are their names [21]. It is the dendrites' task to pick up electrical impulses. The electric signal is processed by soma. The resulting information is sent via the axon to the dendritic terminals of the following neuron. A neuron's nucleus is its central processing unit. Neural networks are interconnected systems of neurons that allow for the conduction of electrical impulses throughout the brain.

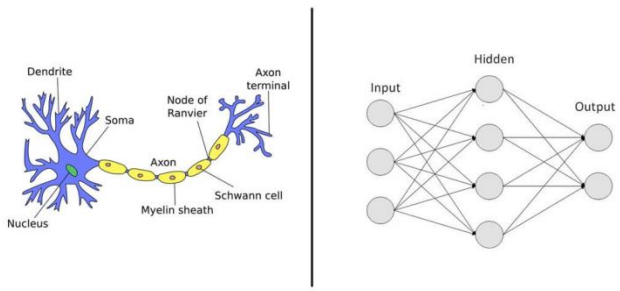

Figure 6: General Structure of neural network

Softmax Function

This concept is expanded upon in Softmax by the inclusion of many social classes. The outer part of neural network models that forecast a multinomial probability distribution uses the soft-max function as the activation function. This indicates that it gives each classes in a multi-class problem decimal probabilities. In other words, the soft-max function is the activation function for multi-class classification problems when class membership is required on more than two class labels.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} \quad e^{z_{j\cdots}}} \ldots (1)$$

σ = softmax

$\vec{z}$ = input vector

e^{z_i} = standard exponential function for input vector

K = number of classes in the multi-class classifier

e^{z_j} = standard exponential function for output vector

e^{z_j} = standard exponential function for output vector

F. Proposed Flowchart

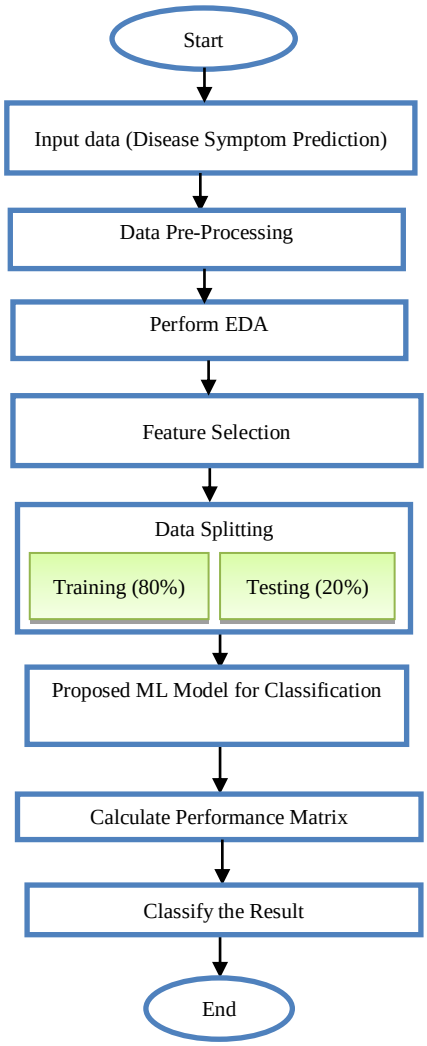

Figure 7: The suggested system's workflow

The following figure 7 shows the proposed flowchart. The first thing that has to be done is to gather all of the pertinent data and then run it through some intelligent data-cleaning procedures. In the next step, the data set is partitioned into two distinct parts: training data and testing data. This step is performed to efficiently create the machine learning model. When the ML model is complete, it is validated by being subjected to many different symptom combinations, and the final output is checked.

IV. RESULT ANALYSIS ANDDISCUSSION

In this part, the outcomes of suggested machine learning models using the disease dataset will be provided. The research was conducted on a machine with Windows 10 that had 16.0GB of physical memory and was powered by a 2.6GHz Intel Core i7-9750H central processing unit. All of the experiments were carried out using the Python programming language, and they made use of the Jupyter simulation environment. Additionally, certain Python libraries were used. This section also includes the performance metrics that are used to figure out how well machine learning models work. Additionally, visualize the input dataset in the form of EDA visulization, simulation, as well as NLP outputs.

A. Performance Paramters

The model's quality is measured using evaluation measures. When diagnosing problems, it's typical to classify them into two categories: those with disease and those without. With this in mind, the goal of such a system is to properly classify data samples and determine which ones belong to which classifications. Although there are other quality evaluation metrics available, some of the most well-known are ROC, f1-score, accuracy, recall, as well as precision. In four ways, those metrics contrast the system's projected classifier performance with the actual classification performance [22].

1) Confusion Matrix

A visual depiction that can be used to explain categorization criteria is the confusion matrix (CM). This matrix contrasts different classes depending on the systems' predictions as well as the actual class of a particular instance of data. It is not a measurable statistic that stands on its own but rather is dependent on the following four components: TP, FP, TN, and FN are outlined in the following paragraphs:

True Negative (TN): True negative results are those for which it can be shown that the model properly predicted the lack of the target class.
True Positives (TP): True positive results are those for which the model can determine that the target class exists.
False Positives (FP): A result is said to be measured as FP when the model incorrectly assesses the existence of a positive class.
False Negative (FN): A result is deemed false if the model incorrectly expects the presence of negative classes.

2) Accuracy

The proportion of data samples that are successfully identified in a classification issue in relation to the overall number of sample data along with forecasts determine the limits on accuracy. We did this by dividing the quantity of correctly identified samples by the sum of TP & TN (the main diagonal of the CM).

$$Accuarcy = \frac{TP + FN}{TP + TF + FP + FN} \ldots\ldots(2)$$

3) Precision

System precision may be calculated by contrasting the true positive (TP) with all positive occurrences (TP + FP). A simple division by the sum of the two components (TP + FP) will do the work.

$$Precision = \frac{TP}{TP + FP} \ldots\ldots(3)$$

4) Recall or Sensitivity

The sensitivity statistic illustrates how many "positive" events were positive. The formula is essentially TP divided by the total of TP + FN.

$$Recall = \frac{TP}{TP + FN} \ldots\ldots(4)$$

5) F-measure

The probability that a positive prediction will turn out to be accurate is shown below. It performs the necessary mathematical operations to calculate the harmonic mean among two values. As a consequence of this, both the FN and FP points of view are taken into consideration. Apply the continuity formula to get an individual user's F1 score:

$$F1 - score = \frac{2(precision \times recall)}{precision + recall} \ldots\ldots(5)$$

B. Simulation Results

The experiments carried out to assess the effectiveness of the suggested model are presented in this section. The experiments withDisease Symptom Prediction data are displayed in various formats, including a confusion matrix, EDA, word clouds and feature correlation ,etc.

EDA is a method of data analysis that locates recurring patterns in the collected information. These patterns consist of peculiar occurrences in the data, known as outliers, as well as paticular characteristics of the data. EDA stands for exploratory data analysis and is a method that employs visual approaches. It can be used to spot patterns and trends as well as to verify assumptions with the aid of statistical summaries & graphical representations. Please share the EDA of Kaggle's input disease dataset on python simulation in the following area.

Figure 8: Symptoms World Cloud

Figure 9: Disease Word Cloud

Figures 8 and 9 make it easy to see the most important keywords or semantic phrases drawn from the whole wordcloud dataset used for the NLP word. A "word cloud" is a graphical depiction of the frequency of individual words. When a term is found

inside the text being analyzed more often, it results in a greater representation of that term in the picture that is created. Word clouds are becoming more popular as a straightforward instrument for determining the primary topic of written content.

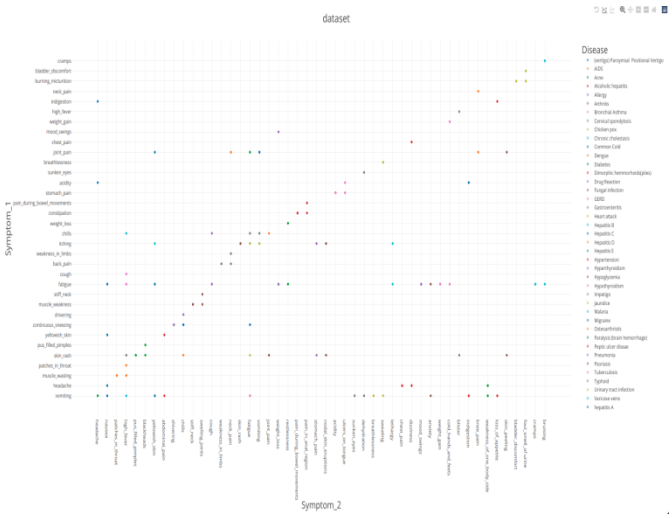

Figure 5: Symptom disease correlation

Figure 5 illustrates the illness correlation of the symptoms occurring in the dataset as well as identifies which symptom is linked to the majority of the diseases. The data collection includes several illnesses and their related symptoms. When a user enters a list of symptoms, the model determines which illness is most likely to be the cause. Using this strategy, we may determine the most probable disease associated with the aforementioned collection of symptoms.

	model	score_val	pred_time_val	fit_time	pred_time_val_marginal	fit_time_marginal	stack_level	can_infer	fit_order
0	LightGBM	1.0	0.003134	2.585495	0.003134	2.585495	1	True	3
1	LightGBMXT	1.0	0.003580	4.222154	0.003580	4.222154	1	True	2
2	LightGBMLarge	1.0	0.003804	7.147746	0.003804	7.147746	1	True	11
3	NeuralNetTorch	1.0	0.004245	3.493430	0.004245	3.493430	1	True	10
4	XGBoost	1.0	0.004704	1.748468	0.004704	1.748468	1	True	9
5	CatBoost	1.0	0.005327	5.460105	0.005327	5.460105	1	True	6
6	WeightedEnsemble_L2	1.0	0.005709	5.893086	0.000382	0.432981	2	True	12
7	ExtraTreesGini	1.0	0.093132	0.624035	0.093132	0.624035	1	True	7
8	NeuralNetFastAI	1.0	0.095586	10.828547	0.095586	10.828547	1	True	1
9	RandomForestEntr	1.0	0.114578	0.686318	0.114578	0.686318	1	True	5
10	ExtraTreesEntr	1.0	0.114604	0.699080	0.114604	0.699080	1	True	8
11	RandomForestGini	1.0	0.115317	1.039083	0.115317	1.039083	1	True	4

Figure 6: Model Score time

The following figure 6 shows the model scoretime on the training period. For this, we implement 12 models that are already described in the methodology section and the above figure also shows the score validation that is 100% and calculates other input variables.

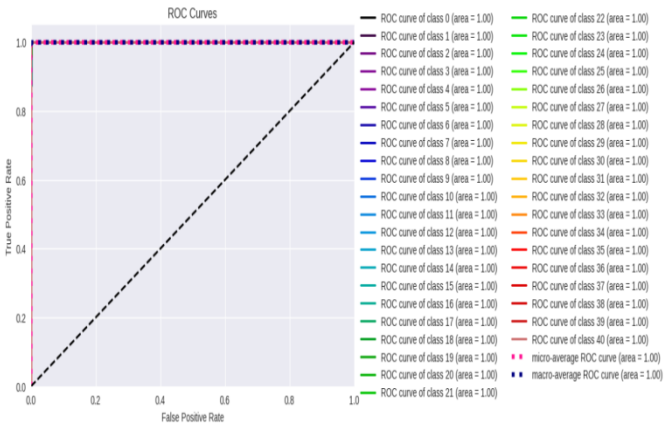

Figure 7: Graph ROC curve

Using ROC curves, as illustrated in figure 7, is a common technique for doing so. The relationship between the classifier's sensitivity as well as its specificity is shown in a ROC curve (1–specifcity). To generate ROC curves, all one has to do is rank the highest scores and change the cutoff point for a two-class issue using a scoring classifier (i.e., a classifier that generates a score for every possible class as well as predicts the class with the biggest score).

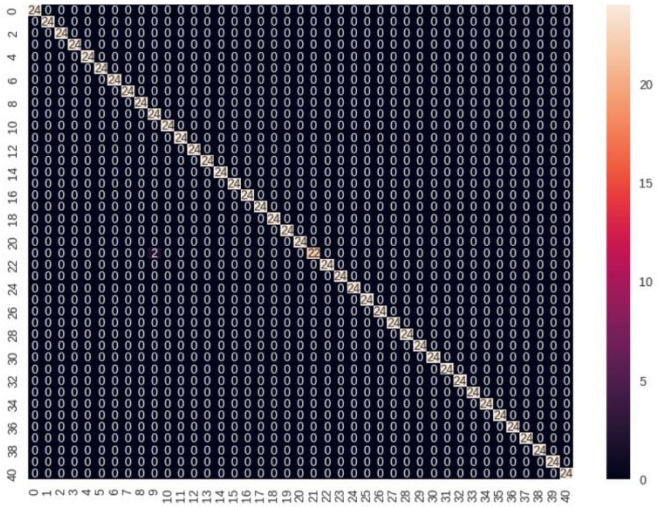

Figure 8: Confusion Matrix

Figure 8 depicts a confusion (or error) matrix, a special kind of table layout that may be used to quickly get a sense of how well a supervised machine learning algorithm is doing. Cases (instances) belonging to a certain illness (class) are represented in the rows of the matrix, while instances of the corresponding diseases (instances) are represented in the columns (class). Both computationally, to generate different performance indicators, and visually, to detect which illnesses (classes) the algorithm confuses, confusion matrices play a crucial role.

```
Diseases: 0  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 1  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 2  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 3  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 4  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 5  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 6  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 7  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 8  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 9  Precision: 0.9230769230769231  Recall: 1.0  Fscore: 0.9600000000000001
Diseases: 10  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 11  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 12  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 13  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 14  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 15  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 16  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 17  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 18  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 19  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 20  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 21  Precision: 1.0  Recall: 0.9166666666666666  Fscore: 0.9565217391304348
Diseases: 22  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 23  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 24  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 25  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 26  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 27  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 28  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 29  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 30  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 31  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 32  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 33  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 34  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 35  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 36  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 37  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 38  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 39  Precision: 1.0  Recall: 1.0  Fscore: 1.0
Diseases: 40  Precision: 1.0  Recall: 1.0  Fscore: 1.0
```

Figure 9: Performance of parameters the input dataset

The following figure 9 shows the performance of parameters the input disease dataset. The high precision, recall ,and f1-score is 100% while class 9 disease precision and recall are 92%, 91% ,and 100% respectively. Similarly for the other dataset classes precision, recall and f1score show their predicted performance.

```
{'accuracy': 0.9979674796747967,
 'balanced_accuracy': 0.9979674796747967,
 'mcc': 0.9979208922891476}
```

Figure 10: Accuracyof proposed model

The following figure 10 shows the performance of accuracy on the input disease dataset using ML models. The accuracy of the proposed model is 99%, and balanced accuracy and MCC are also 99% respectively.

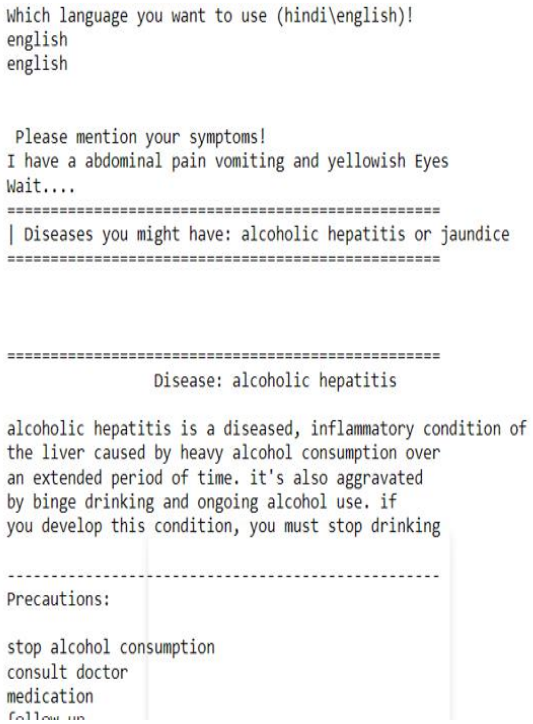

Figure 11: Voice instructions in English

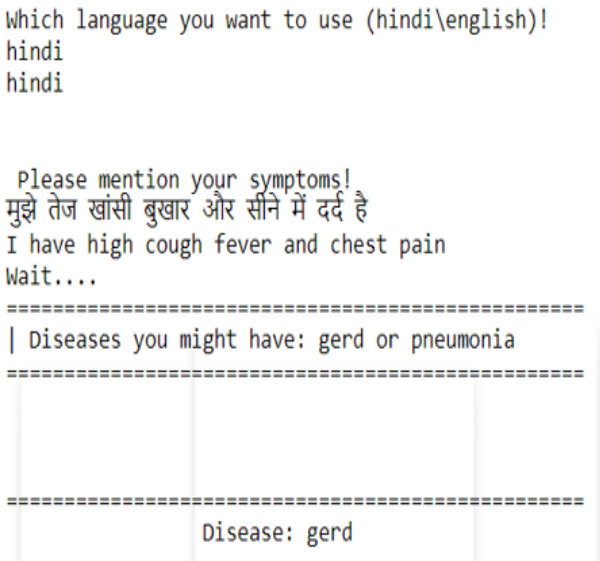

Figure 12: Voice command in Hindi

V. CONCLUSION AND future WORK

The field of machine learning has seen significant advancements over the last decade, and it is now being effectively used in a broad variety of intelligent applications that span a variety of data-related issues. It is essential to analyze the data that already exists to be able to accurately diagnose diseases based on symptoms since the volume of medical data is expanding at an exponential pace. Numerous general disease prediction systems, including LGBM, XGBoost, CatBoost, RF, Extra Tree, & Neural Network, among others, have been created to detect patient data. These systems are based on the natural language processing (NLP) & machine learning (ML) methods. Using this strategy, accurate forecasting of illness and danger may be accomplished in a short period time and for a cost that is quite low. In a nutshell, those that are consistently worried about their health stand to gain from using our approach. Our goal is to establish this system in to make people more aware of the problems affecting their health and to provide them the opportunity to live better lifestyles. We are concentrating on providing our customers with a fast and precise illness prediction based on the symptoms they input, in addition to projecting the severity of the sickness that they may be suffering from. It will give the most effective algorithm as well as consultation with a medical professional. In order to predict illnesses, a number of machine learning algorithms are applied, guaranteeing both speedy and precise predictions. In addition, our system will deliver more accurate disease estimates, as well as initially assumed pictures and text, based on the user's symptoms. As a direct result of this, we are in a position to assert that our system does not include any restrictions since it is open to the participation of everyone.

VI. REFERENCES

[1] S. Tian, W. Yang, J. M. Le Grange, P. Wang, W. Huang, and Z. Ye, "Smart healthcare: making medical care more intelligent," *J. Glob. Health*, 2019, doi: 10.1016/j.glohj.2019.07.001.

[2] B. Zhou, G. Yang, Z. Shi, and S. Ma, "Natural Language Processing for Smart Healthcare," *IEEE Rev. Biomed. Eng.*, pp. 1–17, 2022, doi: 10.1109/rbme.2022.3210270.

[3] T. Young, D. Hazarika, … S. P. C., undefined 2018,and undefined 2018, "Recent trends in deep learning based natural language processing," *Ieeexplore.Ieee.Org*. 2018.

[4] A. Musaev, D. Wang, and C. Pu, "LITMUS: A Multi-Service Composition System for Landslide Detection," *IEEE Trans. Serv. Comput.*, 2015, doi: 10.1109/TSC.2014.2376558.

[5] C. Crema, G. Attardi, D. Sartiano, and A. Redolfi, "Natural language processing in clinical neuroscience and psychiatry: A review," *Front. Psychiatry*, vol. 13, no. September 2022, doi: 10.3389/fpsyt.2022.946387.

[6] J. Prakash and S. Uma, "Disease Identification using Machine Learning and NLP," vol. 1, no. 3, pp. 78–92.

[7] H. Hussain, K. Aswani, M. Gupta, and G. T. Thampi, "Implementation of Disease Prediction Chatbot and Report Analyzer using the Concepts of NLP, Machine Learning and OCR," *Int. Res. J. Eng. Technol.*, vol. 7, no. 4, pp. 1814–1819, 2020, [Online]. Available: www.irjet.net

[8] S. Doan, E. W. Yang, S. Tilak, and M. Torii, "Using Natural Language Processing to Extract Health-Related Causality from Twitter Messages," 2018. doi: 10.1109/ICHI-W.2018.00031.

[9] K. P. Subiksha, "Improvement in analyzing healthcare systems using deep learning architecture," 2018. doi: 10.1109/CCAA.2018.8777545.

[10] M. D. Dascalu, I. C. Paraschiv, B. Nicula, M. Dascalu, S. Trausan-Matu, and A. C. Nuta, "Intelligent Platform for the Analysis of Drug Leaflets Using NLP Techniques," 2019. doi: 10.1109/ROEDUNET.2019.8909606.

[11] X. Yang *et al.*, "A Natural Language Processing Tool to Extract Quantitative Smoking Status from Clinical Narratives," 2020. doi: 10.1109/ICHI48887.2020.9374369.

[12] Z. Yu *et al.*, "Identify Diabetic Retinopathy-related Clinical Concepts Using Transformer-based Natural Language Processing Methods," 2021. doi: 10.1109/ICHI52183.2021.00089.

[13] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using the k-Nearest neighbor algorithm with simplified patient's health parameters," *J. Telecommun. Electron. Comput. Eng.*, 2016.

[14] R. Hossain, S. M. H. Mahmud, M. A. Hossin, S. R. Haider Noori, and H. Jahan, "PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques," 2018. doi: 10.1016/j.procs.2018.05.022.

[15] A. K. Dwivedi, "Analysis of computational intelligence techniques for diabetes mellitus prediction," *Neural Comput. Appl.*, 2018, doi: 10.1007/s00521-017-2969-9.

[16] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," 2017.

[17] Y. Wang, Z. Pan, J. Zheng, L. Qian, and M. Li, "A hybrid ensemble method for pulsar candidate classification," *Astrophys. Space Sci.*, 2019, doi: 10.1007/s10509-019-3602-4.

[18] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *J. Big Data*, 2020, doi: 10.1186/s40537-020-00349-y.

[19] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," *Int. J. Adv. Comput. Sci. Appl.*, 2020, doi: 10.14569/IJACSA.2020.0111190.

[20] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Inf.*, 2020, doi: 10.3390/info11060332.

[21] A. Dey, "Machine Learning Algorithms: A Review," *Int. J. Comput. Sci. Inf. Technol.*, 2016.

[22] J. R. Guerra, "Parkinson ' s Disease Diagnosis : A Machine Learning and Data Mining based Approach," no. February 2019.