Efficient Conversational AI Agent to Improve Rural and Urban Healthcare

Nischita N. J. Mylara Reddy C.

Abstract— Conversational AI agents are software programs which works exactly like humans, they interpret the users and accordingly react to the inputs given by them. These agents are built considering the medical interventions required to improve the overall health of the society. The AI agent designed acts intelligently during the process of the interaction between the humans and itself. It allows the user to use the interface by asking interactive questions then it processes them and responds relatively. Conversational agents are not only web based but they can also be used on other platforms like mobile phone or any other mobile devices. Despite all these a user shall be satisfied if and only if the software is easy to use and obtains the exact results with all of the queries being answered. The main concern with this model is to give that ease to the user to interact with the agent thus solving the queries related to the symptoms suffered by the patients and hence predicting the disease at an early stage by maintaining the accuracy. There are around 100000 diseases in the world according to WHO. Most of their symptoms overlap as well hence by using this agent its possible by it to think insightfully and predict the early symptoms of the disease. In this paper we have designed a user interface and this interacts with the user to take the necessary inputs. This data is fed to the advanced Natural Language Understanding (NLU) to provide the personalized prediction based on the user interaction. The predictions done by the model uses the classification algorithms of Machine Learning. The accuracy of each of these algorithms varies. Therefore instead of considering only one algorithm and hoping it gives the best accuracy, we can use the Ensemble learning method to improve the overall prediction rate. This method gives better predictive indications as it combines many models results thereby improving the overall precision. Here we train our model using various algorithms and ensemble them to get the final results based on the technique of voting. This paper presents a front-end interface for common man using HTML and Angular JS, NLU for text pre-processing using Tensorflow method and ML model as a classifier, for the prediction which uses various machine learning algorithms like SVM, Decision Tree, Random forest etc and combines them all in a majority voting ensemble for balanced results. Therefore this model interacts with any patients be it from the rural or the urban and based on their symptoms predicts and ranks the most probable disease accurately and reliably.

Keywords— Conversational Agent, Artifical Intelligence, SVM, Decision Tree, Random Forest, Ensemble Learning, TensorFlow word embedding

Manuscript published on 30 May 2019.
***** Correspondence Author (s)

Nischita N. J. REVA University, India.
Mylara Reddy C. REVA University, India.

I. INTRODUCTION

Unhealthy behaviours such as smoking, alcohol abuse and lack of physical activities are the sources of cause for disease. Behavioural modifications can lead to a healthy life. But during this process of modification of the lifestyles the person might suffers from some health related issues for which early detection is essential. This involves cost factor and time for the patients to visit the professional doctors[1]. This gap can be bridged by using computer aided systems considering their easier accessibility.

According to WHO most of the deaths in the world are due to non-communicable disease. Amongst them most of them could be detected in prior were avoidable. This requires diagnosis at the primary stage itself and our AI agent can perform this task well.

Technological advancement in natural language processing and artificial intelligence has led to increase in the use of conversational agents. There are few voice activated systems such as Apple's Siri, Amazon Alexa etc[2] but most of these allows only constrained user input. The recent advancement in the machine learning has led to unconstrained natural language input thus leading to complex conversational management along with flexibility. Machine learning techniques are widely used in prediction of the diseases. Medical analysis has become a new trend in the medical sciences. There are huge amount of data generated in the healthcare systems. By using the computational analysis on this clinical data, the agent can be trained to create a medical. intelligence system that can further be used to predict the disease. Thus by developing medical intelligence system, the agent is more patient centric as it already possess the pre-requisite information necessary about the concerned diseases. This in-turn will reduce the cost and saves time for the user. This conversational agent is built considering the diagnosis of the disease at an early stage in a cost effective manner and later bringing awareness amongst the patients to take necessary steps accordingly. Just like the way hospitals will have doctors to enquire the patients about the symptoms they are suffering by cross questioning, similarly here the User Interface along with the NLU & advanced ML algorithms first diagnose the disease using textual conversations and later predicts whether it's a major or a minor disease. Sometimes during this diagnosis its necessary for the doctors to be available at the same time to have experience but in our model since it is trained with different use cases it can handle most of the situations well. For the process of prediction since it uses different techniques like Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT),

Retrieval Number: E10340585S19/2019©BEIESP
DOI:10.35940/ijeat.E1034.0585S19
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

K-nearest Neighbour (KNN), Logistic Regression (LR), Gradient Boosting (GB) and later ensembles them using the concept of majority voting, therefore there is high rate of accuracy obtained.

The remaining section of the paper is organized as follows. Section II discusses the related works followed by section III about the methodology later Section IV discusses results and Section V talks about References.

II. RELATED WORK

This section describes about some of the key researches done in the field of disease diagnosis using the machine algorithms and the ensemble methodology.

Conversational Agents improving every year through various techniques for multiple purposes. Many techniques and methods are employed on them for their advancement including the integration of Natural Language Processing. Also the research says that use of Machine Learning has enhanced the communication of these agents.

Ensemble method is like a meta-algorithm used in various applications due to its accuracy prediction. Considering this feature it's used in the healthcare industry also. Good amount of research has been done in this field.

Najmeh Fayyazifar et al. [3], investigated the use of AdaBoost and Bagging algorithm as classifiers to detect Parkinson's Disease. They found the accuracy rate was 98.28%.

Tahira Mahboob et al. [4] in their research of prediction of coronary heart disease used various methods like SVM, KNN, ANN on 13 attributes for 50 instances. They found the result obtained by each of the methods varied according to the attributes and hence concluded that a combination of all the algorithms using the concept of voting gave them accurate results according to different scenario. The same results were plotted in confusion matrix also to check for the accuracy.

Liangyuan Li et al.[5] used peritoneal dialysis database for analysis and later used the ensemble methodology to predict peritonitis in a cost efficient manner. They conducted the research on real cases to prove the effectiveness of the ensemble method.

Madhuri Gupta et al.[6]. Proposed the model for detection of breast cancer detection using ensemble method. They used feature extraction technique and voting technique to achieve improved prediction.

Qiao Pan et al.[7] used Bagging, Random forest and Bagging methods to predict thyroid disease. They used the ensemble method on different datasets and attributes like age, gender, T3, T4, TSH etc. and found the accuracy to be around 96.16%.

Md. Jamil-Ur Rahman et al.[8] proposed an ensemble Robust Intelligent Heart Disease Prediction System (RIHDPS) using naive Bayes, logistic regression and neural network. This could be used as an assistance to doctor for answering questions.

Prof. Dhomse Kanchan B. et al. [9], conducted study on PCA to find the minimum number of attributes required to enhance the disease prediction using machine learning algorithms. They used SVM and decision tree for heart disease and diabetes and found SVM to work better with few attributes.

Different classifiers offer contradictory results and yields unsatisfactory information. It would be helpful to combine multiple classifiers decision and conclude the valid results as sometimes there is a necessity to predict results for dirty data as well.

Likewise there were many works in the field of ensemble methodology for disease prediction which gave efficient results but there was no single known generic model to predict all the disease for a common person. There is a need for easy user interface with good NLU, a generic disease prediction system using advanced machine learning technique for good precision. In our works we try to fill this gap and with the use of Tensorflow word embedding the classifier will get a better input for the processing and thus generates reliable results.

III. METHODOLOGY

The objective of this study is to promote good health to the people by helping them diagnose the disease according to the symptoms faced by them. Early detection could encourage them to take necessary action and consult specialists for further treatment of their disease. This helps them to prevent the onset of some chronic disease.

- Development of a user friendly interactive text-to-text user interface front end.
- Use of good word embedding methodology so that it can extract the meaning from the text which enables the NLU to predict the contextual meaning of the text given by the user.
- Designing a classifier algorithm that predicts the disease reliably, this is done by the ensemble methodology.

At a high level, the system consists of a front end User Interface (UI) for the patients to act as an interaction module. This is constructed using HTML and Angular JS. Angular JS is a structural framework for dynamic web applications that uses HTML as a template. The biggest advantage of using this is due to its data binding capability. Moreover its an open source web application framework.

This module communicates with the NLU at the backend. It's necessary for the machine to understand the users language to give correct analysis. Each user will have different linguistic ability or vocabulary and hence this makes the machine more responsible for understanding their expressions. Therefore, our NLU is trained with TensorFlow word embedding. Compared to other word embedding like Word2Vec or GloVe advantage of this is higher level abstraction, it does not require pre trained word vectors and works on any language. After the exact intent extracted from NLU the same is being fed to the third component that is the Machine learning classifier which produces the desired results and sends the same to the patient displayed on the user Interface.

Our machine learning module uses many classifier algorithms like Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-nearest Neighbour (KNN), Logistic Regression (LR) and Gradient Boosting (GB). The functionalities of each of these described below:

a. NB- This algorithm follows Bayes Theorem, which tells that every feature that is classified will be independent of each other.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

b. RF- This algorithm generates a number of decision trees for different set of attributes and later concludes which tree is best applicable for the given set of problem.
c. DT- This algorithm is a tree like graph constructed based on numerous attributes especially for conditional control statements.
d. SVM- This algorithm is like representation of the attributes as a point in space and bringinf in clear distinction between the set of features.
e. KNN- This algorithm classifies the coordinates into groups that follows a specified attributes and later performs the results evaluation.
f. LR- Its like a binary classification algorithm which establishes a relation with one dependent variable with one independent variable.
g. GB- This is used for both classification and regression especially in generating results for weak classifiers.

As observed in this model we run the inputs on all of the above techniques and later ensemble them using majority voting method to get the desired results.

A brief schematic diagram of the model is shown in Fig 1.

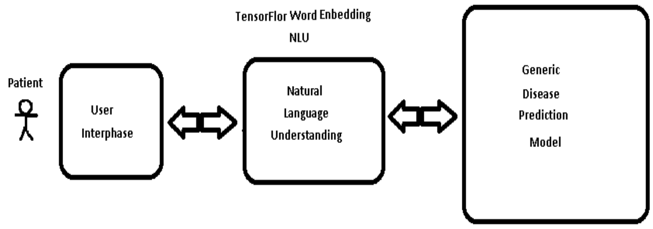

Fig 1: Schematic Diagram

The detailed architecture of the training model is as shown in the Fig 2.

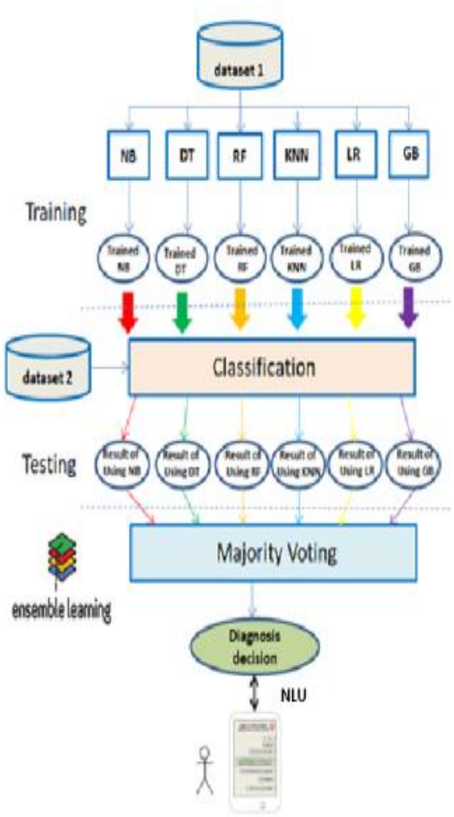

Fig 2: Detailed diagram of the model.

As explained we ensemble the results obtained by each algorithm using majority voting technique. It improves the performance of the given inputs.

As an example in this case we have considered Diabetes prediction. In this case the AI agent interacts with the user by asking a couple of questions like the age, gender, weight, height, symptoms like eating habits etc., to which the user responds accordingly. This data obtained is pre-processed and later is fed to the back-end. At the back end there are two stages first one is the NLU which understands whatever the user has sent and the second is the predictor which predicts whether the patient suffering from diabetes. It uses the ensemble methodology for maintaining the accuracy of its prediction. As a last step the AI agent responds to the user predicting whether the user is diabetic or not.

The detailed architecture of the model is depicted in Fig 3.

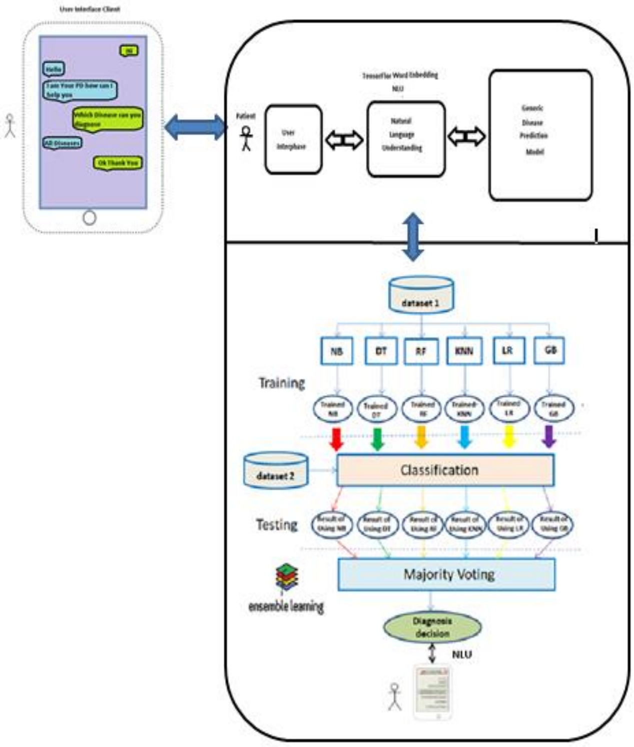

Fig 3: Detailed architecture

IV. RESULTS AND DISCUSSION

The proposed work was performed to test whether the patient is suffering from diabetic as one of the examples. The performance of different classifiers are as shown in the table 1.

Table 1 Performance of different classifiers.

Sl. No.	Classifier	Accuracy (%)
1	Naïve Bayes	52
2	Decision tree	64
3	Random Forest	80
4	K-Nearest neighbour	73
5	Logistic Regression	84
6	Gradient Boost	82
7	Ensemble classifier	84.2

Different algorithms yield different results for the given set of inputs. Based on the attributes the results varies. The comparison of the models accuracy is plotted as shown in Fig 5

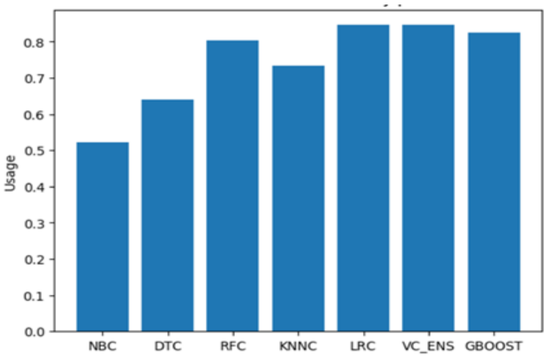

Fig 5: Comparison of Model Accuracy.

It is observed that accuracy is least for Naïve Bayes whereas accuracy is high in Logistic regression and Ensemble classifier. To check for the exact prediction we also plot ROC curve. It shows the performance measurements. ROC curve is shown in Fig 6.

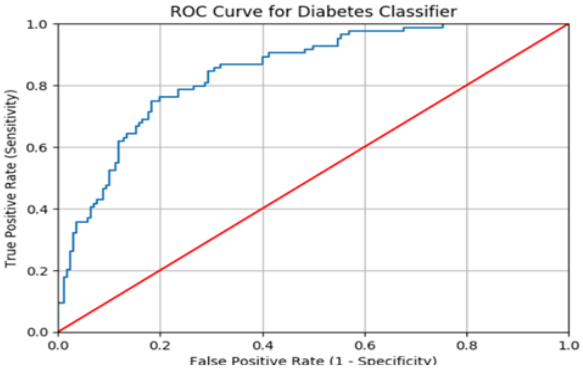

Fig 6: ROC curve for Diabetes

Here sensitivity predicts the exact right values and specificity depicts the exact negative values or the persons not affected. Thus there is higher overall accuracy in the tests being performed.

V. CONCLUSION AND FUTURE SCOPE

To conclude, our end-to-end approach helps in predicting general health issues. Especially with the use of ensemble technique it gives more accurate results. Ensemble classifiers combine all the single classifiers thereby minimizing the error rates that are generated by individual classifiers. This system thus reduces the cost and time for the patients. Upon receiving the advice from this agent the patient can decide about the further steps necessary for his ailments.

As a future scope this model can be improved by using other ensemble algorithms for its accuracy. It can also be used for specific disease prediction as an advanced model or it can also be used by the doctors themselves for prediction of certain diseases. This makes sure that there is no wrong treatment for the patients. The main limitation of this model is to deal with insufficient data. For this the patient can take the help of knowledgeable personnel and provide the information if not the model can suggest other options like visiting the doctors immediately. Better ensemble algorithm can be used for this purpose.

REFERENCES

1. Haolin Wang and Qingpeng Zhang, Mary Ip and Joseph Tak Fai Lau,,” *Conversational Agents for Health Management and Interventions* “, IEEE Computer Society, 2018
2. Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, andEnrico Coiera *“Conversational agents in healthcare: a systematic review*” Journal of the American Medical Informatics Association, 25(9), 2018, 1248–1258
3. Najmeh Fayyazifar and Najmeh Samadiani *“Parkinson's Disease Detection Using Ensemble Techniques and Genetic Algorithm*”, Artificial Intelligence and Signal Processing (AISP), IEEE. 2017
4. Tahira Mahboob, Rida Irfan, Bazelah Ghaffar *“Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics*”, IEEE Conference. 2017
5. Liangyuan Li, Mei Chen, Hanhu Wang, Wei Chen, Zhiyong Guo, ” *A Cost Sensitive Ensemble Method for Medical Prediction*” , First International Workshop on Database Technology and Applications IEEE Computer Society. 2009
6. Madhuri Gupta, Bharath Gupta, “*An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP*)”, Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India.
7. Qiao Pan,Yuanyuan Zhang, Min Zuo, Lan Xiang, Dehua Chen,” *Improved Ensemble Classification Method of Thyroid Disease Based on Random Forest* “,8th International Conference on Information Technology in Medicine and Education. 2016
8. Md. Jamil-Ur Rahman, Rafi Ibn Sultan, Firoz Mahmud, Ashadullah Shawon and Afsana Khan, “*Ensemble of Multiple Models For Robust Intelligent Heart Disease Prediction System*”, 4th International Conference on Electrical Engineering and Information & Communication Technology. 2018
9. Prof. Dhomse Kanchan B. Mr. Mahale Kishor M. *,” Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis*”, International Conference on Global Trends in Signal Processing, Information Computing and Communication. 2016
10. Dinesh Kumar G, Arumugaraj K, Santhosh Kumar D, Mareeswari V , “*Prediction of Cardiovascular Disease Using Machine Learning Algorithms*”, Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.
11. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean “*Distributed representations of words and phrases and their compositionality*,” in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc.,. 2013
12. J. Pennington, R. Socher, and C. D. Manning, “*Glove: Global vectors for word representation*,” in Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. 2014

Retrieval Number: E10340585S19/2019©BEIESP
DOI:10.35940/ijeat.E1034.0585S19
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)