Medical Search Engine

Palak Arora , Shresth Singh ,Megha Rathi
Computer Science Department, Jaypee Institute of Information Technology,Noida ,Sector 62,
Uttar Pradesh , India

Abstract

Data analysis is often described as the process of discovering patterns, correlations, trends or relationships by searching through a large amount of data stored in repositories, databases, and data warehouses.This project helps in identifying a disease based upon the symptoms provided by the user and then according to the predicted disease the project suggest measures for maintaining normal health by showing the prevention , management ,symptomatic , antibiotics and antiviral measures to the user. In this project Artificial Neural Network was used to predict from the dataset of patients at New York Presbyterian Hospital.The Web Scraper is used to show the treatment and prevention information.Results have been shown using an Android application.The data is further visualised on tableau to find out the most prominent symptom that is causing maximum number of diseases in the used dataset.The app also have an additional feature i.e., the user can calculate there Heart Rate in Bpm(beats per minute) in real time .

Keywords: Android Application ;Data Analytics; Diseases; Predictions;ANN;Tableau;Web Scraper;TensorFlow Lite ;deep Learning ;Symptoms.

Introduction

Data Analytics, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies.Having health issues has become at most common in today's world. So for every individual it is important to take a precautionary measure to check if the person has any chances of getting any disease by knowing the symptoms that they may be having. For this purpose we use deep learning techniques to predict if a person has a disease or not , additively telling them the precautions and treatments. [Kavitha et al., 2015].It is attractive as the results are obtained through an android application installed in mobile device.

Data analytics tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. [Shofi et al.,2017] The automated, prospective analyses move beyond the analyses of past events provided by retrospective tools typical of decision support systems. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

The current technology has been used to improve the quality of human life in different ways. One of the areas that most benefit is medicine. However, despite the great progress that the field has made, problems such as the full coverage of the general population, which is presented particularly in developing countries, still remain unresolved . Although the technology has contributed substantially in the way health services are provided, they have not been able to eliminate some problems such as the barriers that a patient has to access the medical service,wait times or the management of all cases presented in the emergency department. One of the solutions that technology proposes is the use of automatic learning methods, that is, applications computing that are trained with data and are capable to perform the same activities as a doctor. [Ramesh et al. ,2016] This type of systems has been proposed since the eighties using different methods such as the use of artificial intelligence. Our project is focused on reducing the time it takes to find the ideal specialist in organisations that provide medical services, with the purpose of focusing the valuable time of the specialists to determine and solve the affliction, leaving the first phase of identification or diagnosis to a computer system. [Khan et al.,2015]The system developed by us consists of an Android application which asks the user to enter the symptoms they are experiencing and then sends that data to our web API which predicts the disease based on the symptoms and also returns possible treatments and care tips for the user to employ against the disease.

Literature Survey

[Shankar et al.,2015] , Method to predict diseases and their cure time based on symptoms reported and severity of these symptoms.They do this by assigning different coefficients to each symptom of a disease, and filtering the dataset with the severity score assigned to each symptom by the user. The diseases are identified based on a numerical value calculated in the fashion mentioned above. For predicting the cure time of a disease, they use reinforcement learning. Their algorithm takes into account the similarity between the condition of the current user and other users who have suffered from the same disease, and uses the similarity scores as weights in prediction of cure time. They also predict the current medical condition of user relative to people

who have suffered from same disease.The dataset used was created by collecting medical information about students in their college. They claim that their approach is better than other proposed approaches as they ask for symptoms and severity rating for each symptom which is similar to the interaction between a doctor and a patient.

[Piñango & Dorado,2014] , Their paper address the problem of prediction of diseases based on specific symptoms in order to improve medical attention given to patients.They propose a flexible Bayesian framework for modelling symptom association with disease in population-based studies. They employ a Bayesian probabilistic model to describe the correlation between specific symptoms such as fever and its cause. The effectiveness of the model is tested with a similarity based measure and training data.The dataset was collected from the openMRS free software project which contained anonymised data of 5000 people and 5,00,000 observations.They claimed to get high accuracy with this dataset.

[Shofi et al.,2017] , Propose to build an Android based application which asks the user some questions and then predicts their diseases based on their answers. They used a forward chaining inference trained on a dataset extracted from a book "Doctor in Your Home" by Dr. Tony Smith. They claim that predictions of their model were highly accurate as validated by experts.

[Palaniappan & Awang,2008] ,Their paper discuss the efficiency of data mining techniques namely: Naive Bayes Classifier , Decision Trees and Neural Networks in predicting the occurrence of a heart disease in a patient using categorical data from patients medical history. The dataset was extracted from the Cleveland Heart Disease database and all numerical data was converted to categorical data. The results show that the neural network and the naive Bayes classifier were able to predict the occurrence of heart disease more accurately than the decision tree.

[Ramesh et al. ,2016] ,Their paper summarise the role of Big Data Analysis in healthcare and the various shortcomings of traditional machine learning algorithms. They argue that the best way to extract useful information from big data is to use a fusion of two or more data mining techniques as they give the best possible result in least amount of time.

[Chen et al.,2017] ,Their paper propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospitals. Their model was tested on a local disease of cerebral infarction which yielded a an accuracy of 94.8 percent.

[Prabakaran & Kannadasan,2018] ,Their paper propose a reliable multi process method combining decision tree techniques and clustering to build a cardiac arrest risk prediction system.

[Gavhane et al.,2018] ,Propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. They used the Cleveland database from UCI library to collect their data set and trained a multi-layered perceptron on it which yielded a high accuracy in predicting the heart disease.

[Ryman et al.,2018], Proposed to identify factors influencing age at symptom onset and disease course in autosomal dominant Alzheimer disease (ADAD), and develop evidence-based criteria for predicting symptom onset in ADAD.Significant proportions of the observed variance in age at symptom onset in ADAD can be explained by family history and mutation type, providing empirical support for use of these data to estimate onset in clinical research.

[Tsanas et al.,2014] ,Proposed to do analysis Nonlinear speech algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity.

[Khan et al. ,2015],The purpose of this study was to develop a method of classifying cancers to specific diagnostic categories based on their gene expression signatures using artificial neural networks (ANNs). We trained the ANNs using the small, round blue-cell tumors (SRBCTs) as a model. These cancers belong to four distinct diagnostic categories and often present diagnostic dilemmas in clinical practice. The ANNs correctly classified all samples and identified the genes most relevant to the classification.

[Jack V. Tu,2017] ,Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes.An overview of the features of neural networks and logistic regression is presented, and the advantages and disadvantages of using this modelling technique are discussed.

Methodology

Sometimes people feel dubious about the things that happened to their physical health and they are not sure what disease affects them. Then, usually, they end up ignoring and underestimating the symptoms they are experiencing. Most people believe that minor illnesses, such as a cold or diarrhoea, do not require special examination or complex treatment. The problem is that a minor

illness could be an indicator of a serious illness. Therefore, one must know whether or not they show the symptoms of a harmful disease and what exactly they should do about it. To make it easier for someone to diagnose the symptoms they are experiencing, in this process we will build an Android-based application which uses a neural network trained on disease - symptom database to predict the disease afflicting the user, that can provide solutions for what they should do when they experience these symptoms, whether they can do their own treatment at home or they should be reviewed immediately by a doctor. By using this application, the diagnosis results of the diseases can be detected through the consultation process or answering the questions provided by the system quickly and effectively with an Android smartphone as a means.

Dataset Description

Dataset is a mixed dataset composed of 1867 instance with 404 symptoms like abdominal pain , cough ,fast_heart_rate etc. containing Binary data and 132 diseases like Diabetes ,Malaria etc containing categorical data . The symptoms that are cause for a disease are marked as '1' and those who are not are marked as '0' for a disease. Appropriate data cleaning and preprocessing techniques are used to convert this data set into 4962 instances with 132 symptoms and 44 diseases , for less redundancy and better prediction by deep learning model.

Approach

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome.The prediction algorithm that we used is Deep learning based Artificial Neural Network.

- **Artificial Neural Network**

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output. ANNs are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modelled or patterns are found. An ANN has several advantages but one of the most recognised of these is the fact that it can actually learn from observing data sets. In this way, ANN is used as a random function approximation tool. These types

of tools help estimate the most cost-effective and ideal methods for arriving at solutions while defining computing functions or distributions. ANN takes data samples rather than entire datasets to arrive at solutions, which saves both time and money. ANNs are considered fairly simple mathematical models to enhance existing data analysis technologies.

ANNs can have many layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to hidden layers, which in turn sends data to the output neurons of the final layer.

- **System Architecture**

The system architecture describes the flow of the project work. The first step in the process is the collection of data needed for the work. Here the dataset used is the data of patients at New York Presbyterian Hospital admitted during 2004, which is collected in the first step. The next step in the process is preprocessing of the data. Here we convert the raw data into understandable format. Now the preprocessed data is classified into a ANN model to predict the person's disease. The user enters the details to know his results for the test into an android app installed in his mobile device. The attributes entered by the user is compared with the pre-trained ANN model and the results are generated.Further generating the treatment and prevention of the disease.

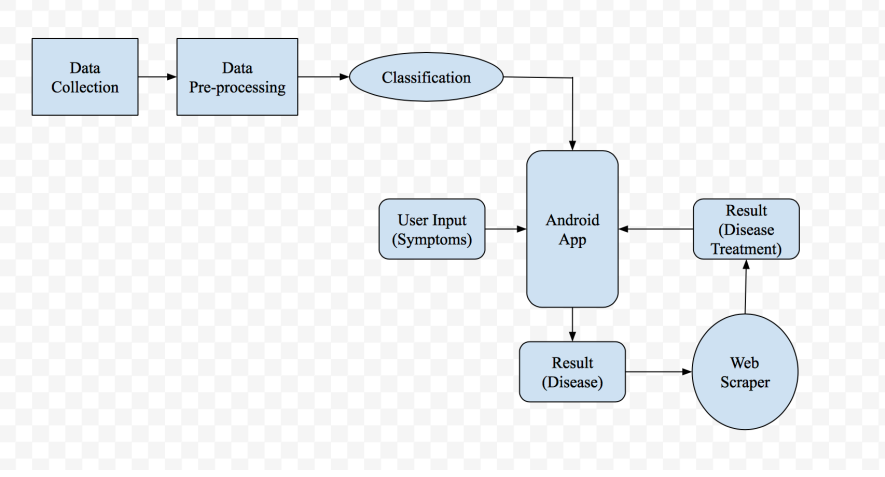

Figure 1 : System Architecture of proposed solution

- **Proposed Solution**

The dataset has been trained with an 4- layer dense ANN using keras classifier with the input layer having 132 input symptoms.Three hidden layers containing 65,33 and 33 neurons from the first layer respectively.The activation function used in all the layers starting from the first layer is relu, tanh,relu and softmax respectively.The output class i.e. prognosis containing 44 classes of categorical type was converted into binary format by using OneHotEncoder , for better prediction of results to remove redundancy.The ANN training is done with batch size of 25 and 500 epochs.The accuracy obtained is 0.8769 i.e. approximately 88%

- Overfitting :

Keeping irrelevant attributes in your dataset can result in overfitting.
It is important to remove redundant and irrelevant attributes from your dataset before evaluating algorithms. This task should be tackled in the Prepare Data step of the applied machine learning process.In our project we have tackled this problem by using Dropouts in all the layers.

The trained keras model is then converted to 'model.h5' file which is further converted to TensorFlow Lite file 'model.tflite' which is easily compatible with Android application.The 'model.tflite' is then added in Assets directory of the Android Studio , android application is made with basic spinners taking symptoms information from the user and then passing it into the model then the predicted disease is displayed on the App along with the probability of the actual occurrence of the disease .These predicted disease is then passed to the Web Scraper which is .ipynb made with with Beautiful Soup in python file hosted on pythonanywhere.The data is scraped from it are the treatment ,management or prevention measured and is displayed on the app as well.

- **Validating Model's Effectiveness**

The measure used for validating the prediction results from the model is "Classification Confusion Matrix".It displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model. The diagonal values show correct predictions.In this example accuracy_score a function of sklearn library is used to find out the value of confusion matrix parameter.The accuracy_score function computes the accuracy, either the fraction (default) or the count (normalize=False) of correct predictions. In multilabel classification, the function returns the subset accuracy.The accuracy_ score comes out to be 0.96 ,i.e. 96% of the prediction are actually correct.

- **Heart Rate Calculator**

Heart Rate Calculator is an extra feature added in the android app that uses the camera and flash light of the user's phone to calculate heart rate by calculating amount of red in preview frames and calculates average.

Working:

- Gets the amount of red in each preview frame.
- Decodes and gets the average amount of red component in image.
- Calculates the rolling average.
- Takes data in chunks of 10 seconds.
- Currently using textureview instead of surfaceview for removing preview of camera from user's eyes.The user can also see for the condition of this heart i.e. where the heart is working Good , Average ,Poor etc. based upon the out coming Heart Rate by seeing through the reference guide provided in the app.

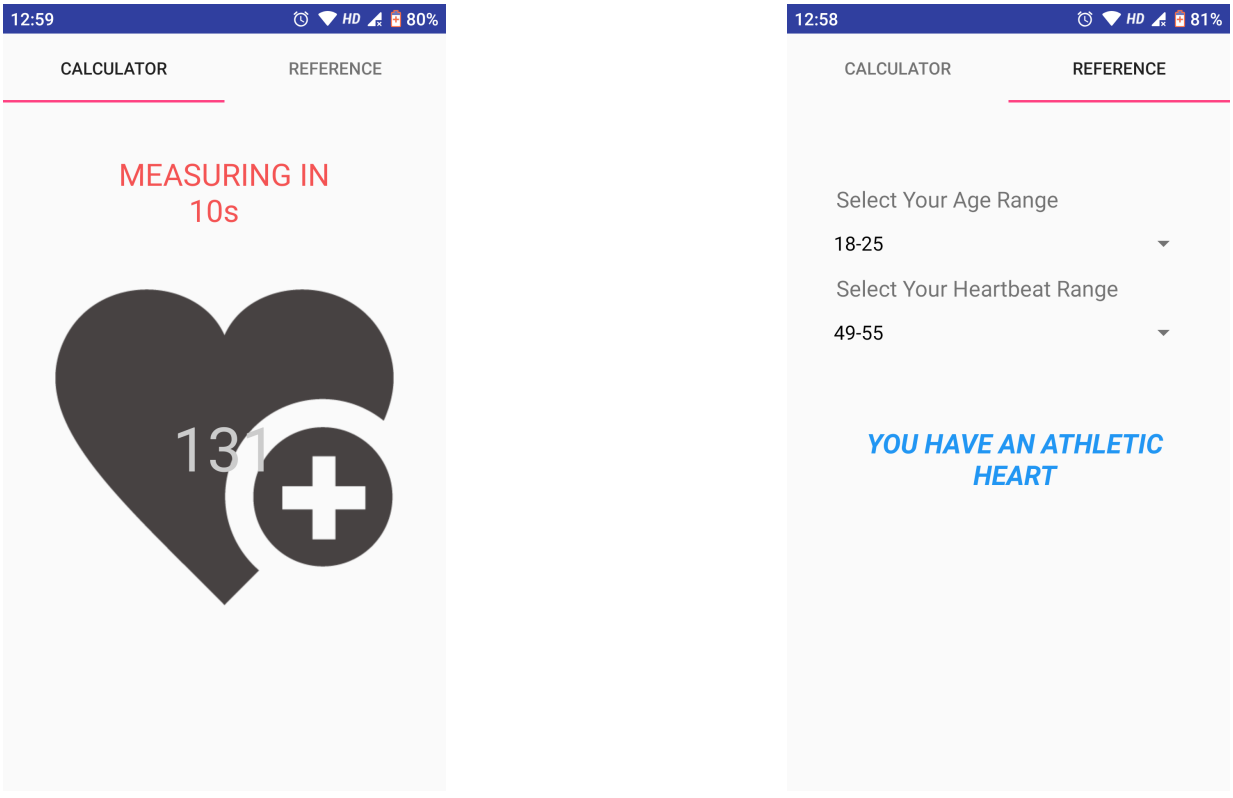

Figure 2 : Functioning of Real-Time Heart Rate Calculator

Result & Analysis

The output of the system will give a prediction result of the disease which the person may deal with.The system gives an idea about the heart status of the user by calculating there Heart rate at real-time.If the person is prone to have a disease then the result obtained can be used to showcase the Treatment and the Prevention measures to prevent or cure the disease. The statistics of the results obtained after the creation of the classifier is shown in the following table :

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 65)	8645
dropout_1 (Dropout)	(None, 65)	0
dense_2 (Dense)	(None, 33)	2178
dropout_2 (Dropout)	(None, 33)	0
dense_3 (Dense)	(None, 33)	1122
dropout_3 (Dropout)	(None, 33)	0
dense_4 (Dense)	(None, 44)	1496

Total params: 13,441
Trainable params: 13,441
Non-trainable params: 0

Figure 3 : Summary of Classifier after model training

```
In [19]: data = pd.DataFrame(l,columns=['Disease','Treatment'])

In [20]: data
```

Out[20]:

	Disease	Treatment
0	adenocarcinoma	<h2>Brain Cancer Treatment Overview</h2><p>Tre...
1	adhesion	<h2>Adhesions Treatment - Self-Care at Home</h...
2	bronchitis	<h2>What Are the Treatments?</h2><p>Most of th...
3	carcinoma	<h2>Treatments</h2><p>Squamous cell carcinoma ...
4	confusion	<h2>Treatment for Sudden Confusion</h2><p>Doct...
5	degenerative polyarthritis	<h2>How Is Arthritis Treated?</h2><p>Treatment...
6	endocarditis	<h2>How Is Hypertrophic Cardiomyopathy Treated...
7	exanthema	
8	hyperbilirubinemia	<h2>How Is It Treated?</h2><p>In adults, jaund...
9	adenocarcinoma	<h2>Brain Cancer Treatment Overview</h2><p>Tre...
10	adhesion	<h2>Adhesions Treatment - Self-Care at Home</h...

Figure 4 : Result of Web Scraper for showing treatment and prevention

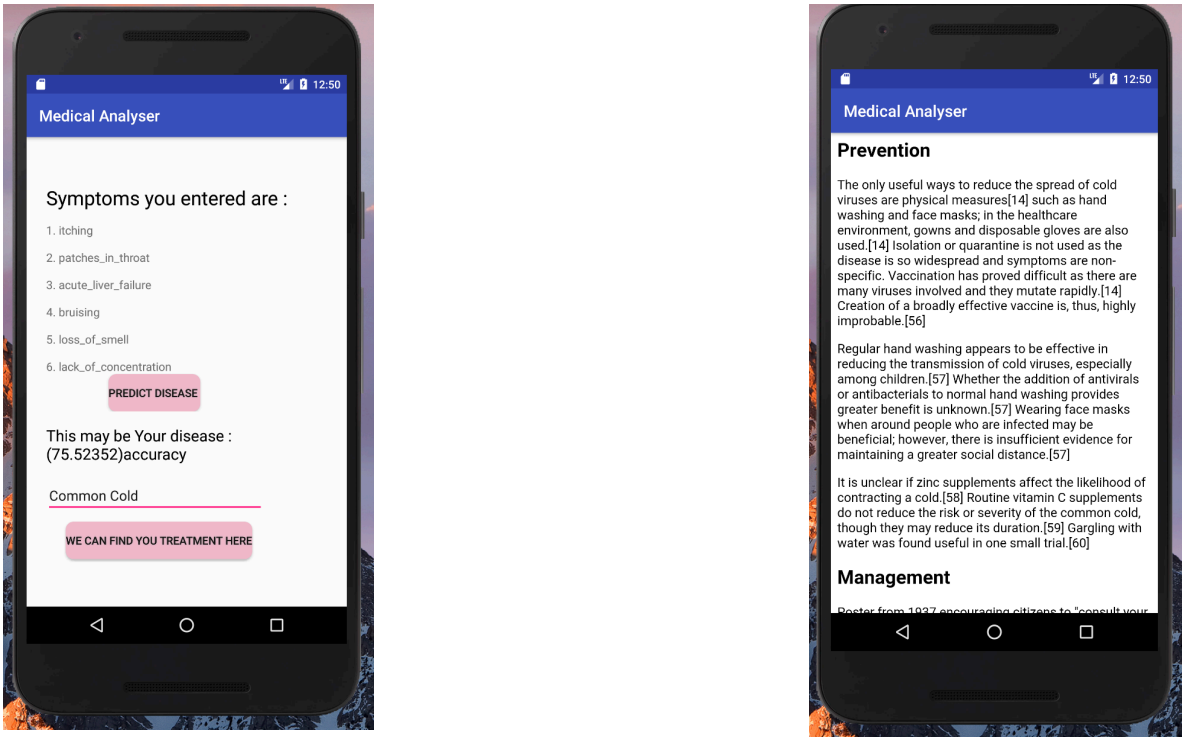

Figure 5: Working of Android Application

Data Visualisation

The most prominent symptom that is the cause of maximum number of diseases in our dataset is found out by using Tableau .By plotting sum of occurrences of ’1’ for each symptom on X-axis and all the symptoms on Y-axis . Fatigue is found out to be the most prominent symptom.

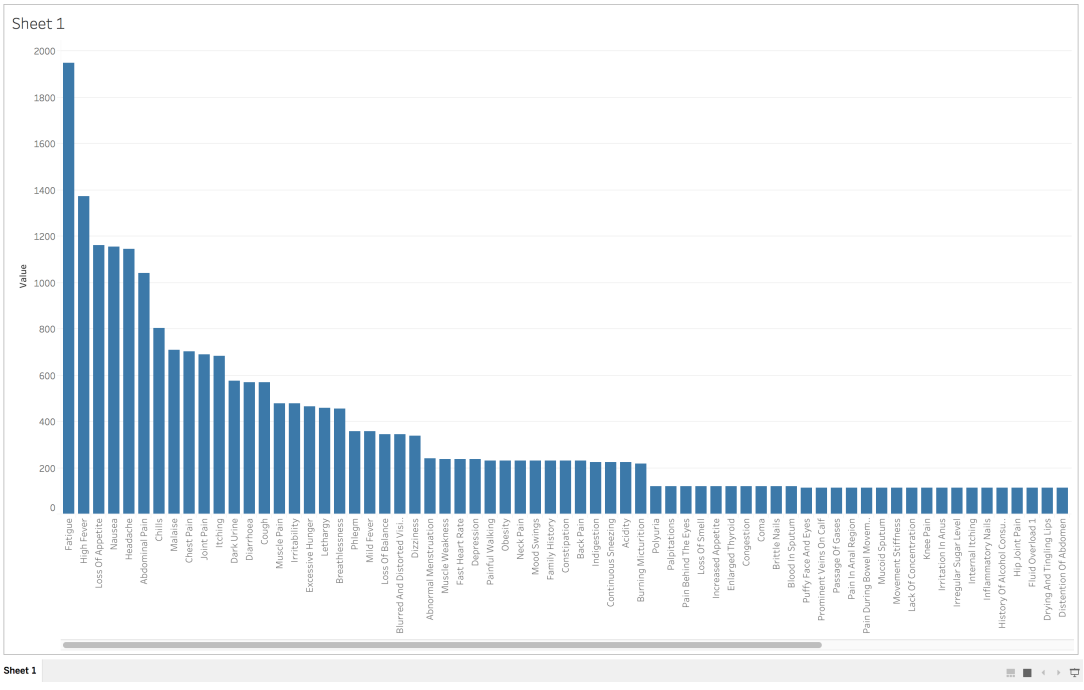

Figure 6 : Histogram, to show the Most prominent symptom in the dataset (Fatigue)

Conclusion

The Medical Search engine System using Machine learning algorithm,viz. ANN provides its users with a prediction result that gives the state of a user leading to any disease. Due to the recent advancements in technology, the machine learning algorithms are evolved a lot and hence we use Artificial Neural Network (ANN) i.e. a kind of Multi Layered Perceptron (MLP) in the proposed system because of its efficiency and accuracy. Also, the algorithm gives the nearby reliable output based on the input provided by the users. If the number of people using the system increases, then the awareness about their current heart status as well as probability of having a disease will be known and the rate of people dying due to the same will reduce eventually.

Future Work

The similar prediction systems can be built that could include various other chronic or fatal diseases ,with the help of recent technologies like machine learning, fuzzy logics, image processing and many others. Also, new algorithms can be proposed to achieve more accuracy and reliability. The Big Data Technology like Hadoop can be used to store huge chunks of data of all the users worldwide and to manage the data or reports of the user; technologies like Cloud Computing can be made use of. Along with this the additional Heart Rate measurement feature that provide heart rate at real time can be added up in the dataset to calculate results based upon the heart rate as well.

References

Sellappan Palaniappan, Rafiah Awang. “Intelligent Heart Disease Prediction System Using Data Mining Techniques,”London,2016 IEEE

Min Chen.,Lu Wang,”Disease Prediction by Machine Learning Over Big Data From Healthcare Community” ,Digital Object Identifier,accepted April 5, 2017

Mani Shankar, Mayank Pahadia, Divyang Srivastava, Ashwin T S, G. Ram Mohana Reddy, Department of Information Technology National Institute of Technology Karnataka, “A Novel Method for Disease Recognition and Cure , 2015 IEEE

Imam M Shofi,Luh Kesuma Wardhani,Ghina Anisa,”Android Application for Diagnosing General Symptoms of Disease Using Forward Chaining Method”.Universitas Islam Negeri Syarif Hidayatullah Jakarta,2017

Aditi Gavhane,Aditi Gavhane,Gouthami Kokkula,Prof. Kailas Devadkar (PhD),”Prediction of Heart Disease Using Machine Learning”,2nd International conference on Electronics, Communication and Aerospace Technology IEEE Conference,2018

Prabakaran.N and Kannadasan.R Scope,”Prediction of Cardiac Disease Based on Patient’s Symptoms”, VIT University, Vellore, India,2018

V.Krishna Priya,A.Monika,P.Kavitha,"Android Application to Predict and Suggest Measures for Diabetes Using DM Techniques",Rajalakshmi Engineering College 2, Flowers road 4th l lane, Purasawalkam,ch-84 Chennai, India,2015

Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE

Shankar, M., Pahadia, M., Srivastava, D., Ashwin, T. S., & Reddy, G. R. M. (2015, May). A Novel Method for Disease Recognition and Cure Time Prediction Based on Symptoms. In 2015 Second International Conference on Advances in Computing and Communication Engineering (pp. 679-682). IEEE.

Ramesh, D., Suraj, P., & Saini, L. (2016, January). Big data analytics in healthcare: A survey approach. In 2016 International Conference on Microelectronics, Computing and Communications (MicroCom) (pp. 1-6). IEEE.

Shofi, I. M., Wardhani, L. K., & Anisa, G. (2016, April). Android application for diagnosing general symptoms of disease using forward chaining method. In 2016 4th International Conference on Cyber and IT Service Management (pp. 1-7). IEEE.

Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of Heart Disease Using Machine Learning. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1275-1278). IEEE.Prabakaran, N., & Kannadasan, R. (2018, April). Prediction of Cardiac Disease Based on Patient's Symptoms. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 794-799). IEEE.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, 8869-8879.

Pinango, A., & Dorado, R. (2014, November). A Bayesian model for disease prediction using symptomatic information. In 2014 IEEE Central America and Panama Convention (CONCAPAN XXXIV) (pp. 1-4). IEEE.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., ... & Meltzer, P. S. (2015). Classification and diagnostic prediction of cancers using gene expression pro filing and artificial neural networks. Nature medicine, 7(6), 673.

Tu, J. V. (2015). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology, 49(11), 1225-1231.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

Palmer, K., Berger, A. K., Monastero, R., Winblad, B., Bäckman, L., & Fratiglioni, L. (2007). Predictors of progression from mild cognitive impairment to Alzheimer disease. Neurology, 68(19), 1596-1602.

Mayberg, H. S., Brannan, S. K., Mahurin, R. K., Jerabek, P. A., Brickman, J. S., Tekell, J. L., ... & Fox, P. T. (1997). Cingulate function in depression: a potential predictor of treatment response. Neuroreport, 8(4), 1057-1061.

Rogers, R., Lombardo, J., Mednieks, Z., & Meike, B. (2015). Android application development: Programming with the Google SDK. O'Reilly Media, Inc..

Meier, R. (2016). Professional Android 4 application development. John Wiley & Sons.

Hansen, L. K., & Salamon, P. (2015). Neural network ensembles. IEEE Transactions on Pattern Analysis & Machine Intelligence, (10), 993-1001.

Krogh, A., & Vedelsby, J. (2015). Neural network ensembles, cross validation, and active learning. In Advances in neural information processing systems (pp. 231-238).