JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.7, July 2020 E-ISSN: 2347-2693

Disease Prediction Using Machine Learning Over Big Data

Ajeesh Babu[1*], Fathima Basheer[2], Jayasanker M[3], Tintu Mariyam Paul[4], Sithu Ubaid[5]

[1,2,3,4,5]Dept. of Computer Science and Engineering, St Thomas College of Engineering and Technology, APJ Abdul Kalam

**Corresponding Author:ajeeshbabu864@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v8i7.1115 | Available online at: www.ijcseonline.org

Received: 04/July/2020, Accepted: 15/July/2020, Published: 31/July/2020

Abstract— Due to big data and progress in biomedical and healthcare communities, accurate study of medical data benefits early disease recognition, patient care and community services. When the quality of medical data is incomplete, the exactness of study is reduced. In the proposed system, our system can take either text or image input symptoms from the user and based on the analysis of the symptoms it displays a result. It provides machine learning algorithms for effective prediction of various disease occurrences in disease-frequent societies. It experiment the altered estimate models over real-life hospital data collected. To overcome the difficulty of incomplete data, it uses a latent factor model to rebuild the missing data. It experiments on various diseases that occur in human being. Using structured and unstructured data from hospital, Random Forest algorithm is used for classification of text datasets. SSD (Single Shot Multi Box Detector) algorithm is used for image processing to analyse various diseases in human being.

***Keywords*—**Big Data, Machine Learning, kaggle, CNN

I. INTRODUCTION

At present, when one suffers from particular disease, then the person has to visit a doctor which is time consuming and costly too. Also if the user is out of reach of doctor and hospitals; it may be difficult for the user as the disease can not be identified. So, if the above process can be completed using an automated program, it can save time as well as money, it could be easier for the patient which can make the process easier. Disease Predictor is a web based application that predicts the disease of the user with respect to the symptoms given by the user. Disease Prediction system has data sets collected from different health related sites. With the help of Disease Predictor the user will be able to know the probability of the disease with the given symptoms. As the use of internet is growing every day, people are always curious to know different new things. People always try to refer to the internet if any problem arises. People have access to internet than hospitals and doctors. People do not have immediate option when they suffer with particular disease. So, this system can be helpful to the people as they have access to internet 24 hours. In this system it has been introduced a text based and an image based disease prediction system. Since it uses big data concepts which make a huge progress in biomedical and healthcare communities, accurate study of medical data, benefits early disease recognition, patient care and community services. It provides machine learning algorithms for effective prediction of various disease occurrences in disease-frequent societies.

II. RELATED WORK

The proposed system predict diseases based on symptoms of the patient. Here k-Nearest Neighbour and Convolution Neural Network algorithms are used. In this, the living habits of the person and the check up information is used for predicting disease[1]. This study proposes the development of Convolution Neural Network (CNN) for an automatic detection system using various deep learning methods for echocardiography.SSD and R-CNN is used. The findings of this study suggested that SSD is suitable for echocardiography as it delivers a good correspondence between prediction, speed and accuracy[2]. Using a fuzzy k-NN algorithm with an artificial immune system diagnosed breast cancerImproves accuracy of k-NN in diagnosing breast cancer by applying genetic algorithm that determines best components for the k-NN algorithmImplementing GA with k-NN resulted in 3% accuracy.Accuracy of the classifier will be calculated using bootstrap sampling method. Major disadvantage is that classification process is quite time consuming[3].CNN-MDRP(Multimodal Disease Risk Prediction) algorithm based prediction is more accurate than previous system using CNN-UDRP(Unimodel Disease Risk Prediction).This technique overcomes the difficulties existed in the previous system such as incomplete and missing data.Here data mining technique is used for disease prediction.[4].This system uses c45 rules and partial tree technique to predict heart disease.Discovered set of rules to predict the risk levels of patients based on given parameter about health condition.Accuracy in testing phase and training phase are 86.3% and 87.3% [5].System extract color features of human nail and stored in RGB form, focuses on image recognition on the basis of human nail color analysis.Color of nail matched using a matcher algorithm.Accuracy : 65% results are correctly matched [6].Clinical data describing the phenotypes and treatment of patients represents an underused data source that has much greater research potential than is currently realized. Mining of electronic health records (EHRs) has the potential for establishing new patient-stratification principles and for revealing unknown disease correlations. Integrating EHR data with genetic data will also give a finer understanding of genotype-phenotype relationships. However, a broad range of ethical, legal and technical reasons currently hinder the systematic deposition of these data in EHRs and their mining. Here, we consider the potential for furthering medical research and clinical care using EHR data and the challenges that must be overcome before this is a reality [7].

III. METHODOLOGY

A. PROPOSED SYSTEM

Now days we see that people are facing various diseases due to many factors in our environment as well as due to our living habits. So the prediction of disease at earlier stage is an important task. In order to solve those problems in existing systems we introduce this proposed system which helps them in self analysis and self-awareness from all these diseases at the earlier stage itself. It helps them to take wanted remedies as earlier. As per the recent applications that are available in the market, the prediction rate is very low for various machine learning algorithms. The big data came into the picture where the database that we have is huge and the same data will be fed inside the software in order to get a very good prediction. Here in this paper, the prediction of the disease is completely on the datasets that is given by the hospital. We will be using big data and also a bit of machine learning algorithm in order to predict the disease. The advantage of cloud can be used for storing huge amount of data. We expect a prediction rate above 90%. In our proposed system we take image and text as input from the users and fed it to our system. We train our system using the datasets downloaded from the Kaggle. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment. Using CNN (Convolution Neural Network) algorithm the training process of text data and image data is done automatically in cloud. The input and model is loaded to the tensor flow memory.

B. SYSTEM ARCHITECTURE

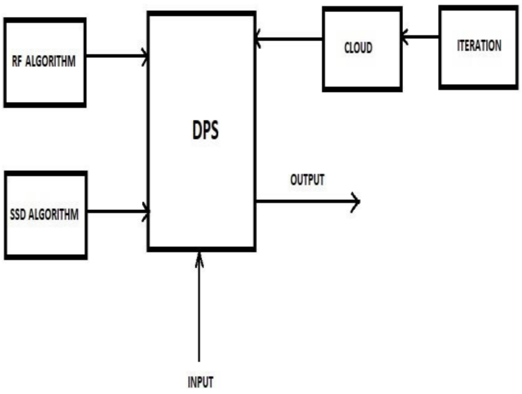


Figure 1: System Architecture

The system architecture includes five steps which are

1. Collect hospital data sample/download the sample from kaggle. Data samples must consists of different kinds of diseases
2. Training of datasets: Here 70% of data samples of text and images datasets are taken for training. Text data samples are trained using classifier algorithm it is done by creating a model. Image data samples are trained by creating a modal called InceptionV3model which is done in deep learning.
3. Testing: Here 30% of data samples of text and image datasets are taken for testing
4. Feature extraction: Using Random forest algorithm and SSD we classify the text and image datasets which recognizes the type of diseases.
5. Prediction: System predicts the types of disease with the input given by the user by using a library function called Flask.

C. TEXTUAL AND IMAGE DATA PROCESSING AND TRAINING

Our system has mainly two sections that is one of the main reasons that make it unique. The two phases include Machine Learning and Deep Learning. The system can take both image and text input from a user one at a time. The system comprises of well-known Machine Learning algorithms, Random Forest and Single Shot Multi box Detection algorithm and deep learning algorithm, Convolution Neural Network (CNN).

These algorithms play a major role in predicting the disease. Textual datasets are dealt by Random Forest algorithm and CNN plays the role of classifying image inputs. The given dataset is converted to data frames. Then they are classified into Training data and test data. Here we used 70% of the data as training data's and 30% for testing. After the classification a model is created using fit () function in python. Training is performed only one time and it is pickled to a file to avoid training every time. The process of writing trained data into a file in python is called pickling.

a. Textual Data Processing and Training

As explained earlier Random Forest plays the major role in processing of the text input given by the user. For training text dataset ML technique is used here. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of over fitting to their training set. A RF takes a number of inputs from the user and yields an output. A library known as Scikit-learn is used in order to do these calculations. Instead of using equations Scikit-learn uses functions. Thus we can provide the best possible result to the users with less error rate.

The sum of the feature's importance value on each tree is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

RFfi$_i$ = the importance of feature i calculated from all trees in the Random Forest model
normfi$_{ij}$ = the normalized feature importance for i in tree j
T = total number of trees

b. Image Data Processing and Training

The image data is processed using Deep Learning technique. For training purpose classifier algorithm is used. In order to train a dataset, first a model should be prepared. Here we are using Inception V3 model. The dataset is collected from Kaggle website. Inception V3 is commonly used in Deep Learning. In our system only two classifications are there, Benign or Malignant. The images in the dataset will pass through these 48 layers which will extract features in order to train the machine. In this project sliding window is used to detect tumours in the images uploaded by user.

To detect a tumour in a test input image, first we have to train a set of images using CNN so that the system will learn how to detect a tumour. For each input region the CNN outputs whether it has a tumour or not. We run sliding window multiple times over the image with different window size, from smaller to larger. Hoping a window size would fit the tumour and allow CNN to detect it.

D. PREDICTION

A web application is developed for prediction, where the user will input all the symptoms and the images of diseases which in turn given to model. The model will predict patients' disorder.

In text based system it uses random forest algorithm for prediction. First it selects random samples from a given dataset. Construct a decision tree for each sample and get a prediction result from each decision tree. Perform a vote for each predicted result. Select the prediction result with the most votes as the final prediction.

In image based system it uses SSD (Single Shot Multi- box Detector) for prediction. SSD discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. At prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape. Additionally, the network combines predictions from multiple feature maps sizes.

A library known as Flask is used for prediction. Flask is a micro web framework written in Python because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

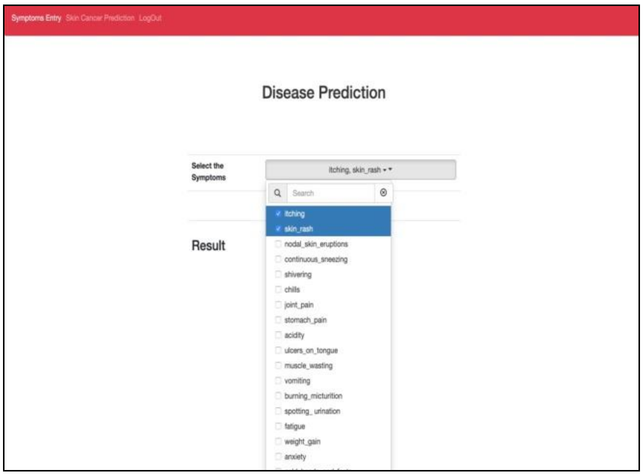
E. CREATION OF USER INTERFACE

We created a platform for the patients to interact to our system.HTML programs are used in order to create the same. Here we used Bootstrap framework which is developed by Twitter to make the website responsive. Cascading Style Sheets are used to style the appearance of the contents inside the web pages. In order to use our system a user needs to register a new account using a username and a password. By providing the details one can enter into their account to check their health condition.

After a user logins to his/her account he/she have 2 options on the top left corner. If the patient wants to input an image then the patient must choose skin cancer prediction otherwise the patient can choose symptoms entry.

IV. RESULTS AND DISCUSSION

Disease Prediction has been already implemented using different techniques like Neural Network, decision tree and Naïve Byes algorithm. From the analysis it was found that Random Forest is more accurate than other techniques to predict diseases with text inputs. And SSD is the algorithm that yields more accurate results by processing image input from the user.



Disease Prediction

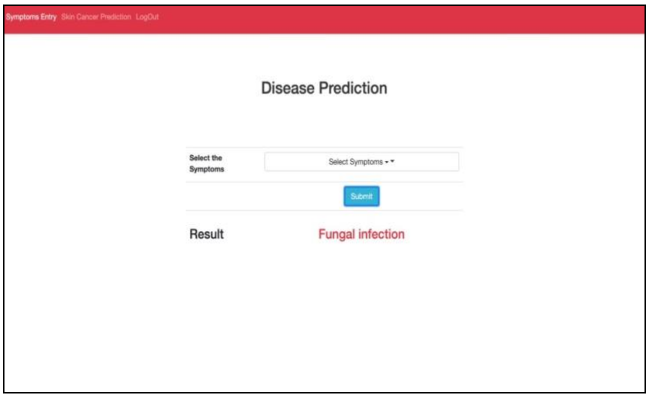
Select the Symptoms: Itching, skin_rash

☑ itching
☑ skin_rash

Result

Figure 2. Selecting Symptoms

After choosing the appropriate symptoms the result will be displayed in the same web page.



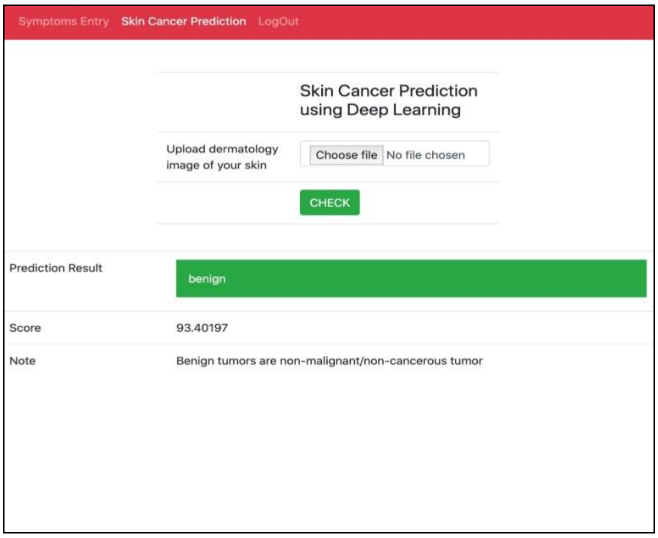
Disease Prediction

Select the Symptoms

Result: Fungal infection

Figure 3. Displaying Results

After the patient selecting the symptoms the corresponding disease will be displayed as in the Figure 4.



Skin Cancer Prediction using Deep Learning

Upload dermatology image of your skin: No file chosen

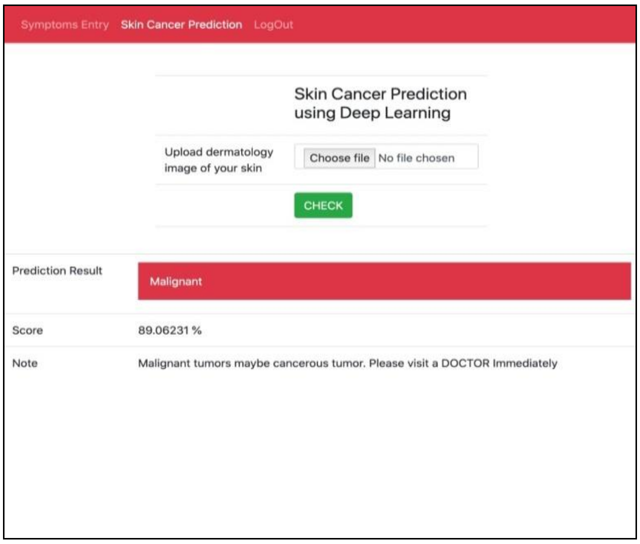
Prediction Result: benign

Score: 93.40197

Note: Benign tumors are non-malignant/non-cancerous tumor

Figure 4. Benign sample screenshot.

If a patient wants to diagnose skin cancer he/she must upload an image of the susceptible part of the skin. Here in this sample screenshot in fig 3.11 the patient have non-cancerous tumour (Benign). The score rate specifies that 93 % the result is correct.



Skin Cancer Prediction using Deep Learning

Upload dermatology image of your skin: No file chosen

Prediction Result: Malignant

Score: 89.06231 %

Note: Malignant tumors maybe cancerous tumor. Please visit a DOCTOR Immediately

Figure 5. Malignant sample screenshot

The result is Malignant (cancerous) and there is 89% chance that the result is correct.

V. CONCLUSION AND FUTURE SCOPE

This paper aims to predict the disease on the basis of the symptoms and image. The paper is designed in such a way that the system takes symptoms and image from the user as input and produces output i.e. predict disease. Prediction accuracy of our project is 95%. Disease Prediction system was successfully implemented using Machine Learning and Deep Learning techniques.

We developed the disease prediction system as a web application. In future this can be developed as a Mobile Application. We can also add a platform which connects users directly to the doctors. Also Disease Predictor does not recommend medications of the disease. And past history of the disease has not been considered.

REFERENCES

[1] M. Chen, Y .Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities”, ,”IEEE Access, vol.5, no. 1, pp. 8869-8879, 2017.

[2] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn,“Wearable 2.0: Enable human-cloud integration in next generation healthcare system,” *IEEE Commun.* , vol. 55, no. 1,pp. 54–61, Jan. 2017.

[3] P. Sharma, R. Rastogi, D.K. Chaturvedi, S. Satya, N. Arora, V. Yadav, S. Chauhan, Analytical comparison of efficacy for electromyography and galvanic skin resistance biofeedback on audio-visual mode for chronic TTH on various attributes, in *Proceedings of the ICCIDA-2018 on 27 and 28th October 2018*. *CCIS Series* (Springer, Gandhi Institute for Technology, Khordha, Bhubaneswar, Odisha, India, 2018)

[4] Shraddha SubhashShirsath “Disease Prediction Using Machine Learning Over Big Data” International Journal of Innovative Research in Science, Vol. 7, Issue 6, June 2018

[5] AnimeshHazra, Arkomita Mukherjee, Amit Gupta, Mukherjee, “Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review”, Research Gate Publications, July 2017, pp.2137-2159.

[6] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, “Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review”, International Journal of Computer Applications, February 2016

[7] AnimeshHazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, “Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review”, Research Gate Publications, July 2017, pp.2137-2159

[8] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, “The ’big data’ revolution in healthcare: Accelerating value and innovation,” 2016.

[9] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.

[10] M. Amiri and G. Armano, "Early diagnosis of heart disease using classification and regression trees," The 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, 2013, pp. 1-4. doi: 10.1109/IJCNN.2013.6707080 .

[11] S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2017, pp. 1-4. doi: 10.1109/EBBT.2017.7956761.

[12] P. Su, J. Yang, Z. Li and Y. Liu, "Mining Actionable Behavioral Rules Based on Decision Tree Classifier," 2017 13th International Conference on Semantics, Knowledge and Grids (SKG), Beijing, 2017, pp. 139-143. doi: 10.1109/SKG.2017.00030 .

[13] D. Bertsimas, J. Dunn and A. Paschalidis, "Regression and classification using optimal decision trees," 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, 2017, pp. 1-4. doi: 10.1109/URTC.2017.8284195Y

Authors Profile

Mr Ajeesh Babu is currently pursuing his Bachelor’s degree in Computer Science and Engineering at St Thomas College of Engineering and Technology, Kerala, India under APJ Abdul Kalam Technological University. He is having interest in the field of Web designing and Python Programming. He is excellent in working in a team and guiding others towards the achievement of objectives.

Ms Fathima Basheer is currently pursuing her Bachelor’s degree in Computer Science and Engineering at St Thomas College of Engineering and Technology, Kerala, India under APJ Abdul Kalam Technological University .She is having interest in the field of digital art and web graphic Designing and she is currently Researching in this area.

Mr Jayasanker M is currently pursuing his Bachelor’s degree in Computer Science and Engineering at St Thomas College of Engineering and Technology, Kerala, India under APJ Abdul Kalam Technological University. He is having interest in the field of cloud computing and cyber security. He is interested in learning about new technologies in the field of information technology.

Ms Tintu Mariyam Paul is pursuing her Bachelor’s degree in Computer Science and Engineering at St Thomas College of Engineering and Technology, Kerala, India under APJ Abdul Kalam Technological University. She is having interest in networking and is currently researching in this area, for obtaining responsible, challenging and rewarding position in the field of desktop support and network maintenance on these areas.

Ms Sithu Ubaid is currently working as an Assistant Professor in Department of Computer Science and Engineering at St. Thomas College of Engineering and Technology, Kerala. Her area of interests are image processing, machine learning, deep learning and artificial intelligence.