International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Implementation of Disease Prediction Chatbot and Report Analyzer using the Concepts of NLP, Machine Learning and OCR

Hiba Hussain[1], Komal Aswani[2], Mahima Gupta[3], Dr. G.T.Thampi[4]

[1,2,3]Student, Information Technology Department, Thadomal Shahani Engineering, Maharashtra, India
[4]Principal, Thadomal Shahani Engineering, Maharashtra, India

---

Abstract - *In recent times, healthcare is becoming more accessible to a wider group of people through the medium of technology. The concepts of artificial intelligence, machine learning and neural networks have provided substantial assistance in the field of healthcare. In today's fast-paced world, people tend to neglect their health which may result in a critical problem. Such a problem can be avoided by using the symptoms driven disease prediction application. Our project focuses on providing the users immediate and accurate prediction of the diseases based on their symptoms along with a detailed analysis of their pathology reports. The disease prediction chatbot is developed using natural language processing and machine learning algorithms. For the prediction of diseases, we have used two classification algorithms namely, Decision tree and KNN (k-nearest neighbors). The performance of these techniques are compared and based on their accuracy, the best model is selected. As per our results, the accuracy of Decision Tree and KNN are 92.6% and 95.74% respectively. This project also looks forward to providing medical consultation on the predicted disease. The pathology report analysis is performed using the concept of Optical Character Recognition (OCR). Tesseract is an open-source recognition engine to perform OCR. The text extracted from the report is used for interpreting the results in an easier way and to provide a graphical analysis of the test results.*

Key Words: **Disease Prediction, Machine Learning, Decision tree, K-nearest Neighbors, Natural Language Processing, Chatbot, WordNet, Python, Optical Character Recognition, Tesseract**

1. INTRODUCTION

Since the past few decades, humans have been tirelessly working day in and day out that they fail to prioritize their health on a regular basis. In the longer run, this problem leads to jeopardizing the quality of life. Nevertheless, with the aid of Artificial Intelligence, we can now provide health care services to individuals at their convenience at reasonable prices. One of the biggest blessings we possess is a healthy body. A healthy body and enhanced quality of life is something each one of us looks up to. The primary focus of this paper is to provide these services to fulfill the above mentioned purpose. It is difficult to imagine our lives without high tech gadgets because they have become an essential part of our lives. Therefore the field of Artificial Intelligence is prospering due to the various applications of it in the research field. Disease prediction is one of the main goals of the researchers based on the facts of big data analysis which in turn improves the accuracy of risk classification based on the data of a large volume. [1]

E-healthcare facilities in general, are a vital resource to developing countries but are often difficult to establish because of the lack of awareness and development of infrastructure. A number of internet users depend on the internet for clearing their healthcare based queries. We have designed a platform for providing online medical services to patients with a goal to provide assistance to healthcare professionals. The user can also seek medical guidance in an easier way and get exposure to various diseases and diagnosis available for it. In order to make communication more effective, we have implemented a chatbot for disease prediction.

Chatbots are the human version of software that is based on AI and uses Natural language processing (NLP) to interpret and accordingly respond to the user. This study proposes the disease prediction chatbot using the concepts of NLP and machine learning algorithms. The prediction is carried out using KNN and Decision tree algorithms. KNN and Decision tree are a few of the most used classification algorithms that are frequently used in disease prediction. It is assisted with the NLP driven chatbot. [2] The wordnet and tokenization concepts of NLP are used. The use of tokenization is to split the given text into a list of words whereas WordNet is a lexical database of dictionary designed for natural language processing. The study also focuses on the use of the Optical Character Recognition tool named Tesseract which is used to extract text from the patient's scanned pathology report. The generated text helps in translating the report in an easier manner by providing a graphical analysis of the test result.

2. LITERATURE SURVEY

Advancement in technology has a far-reaching effect in the field of Healthcare. Machine Learning algorithms have not only helped the doctors but also have provided a first-hand testing set for the patients. Natural Language Processing being a part of AI provides data extraction and assistance in understanding the patient's words in a better way.

A. Disease Prediction

Machine learning uses algorithms that analyze an acceptable range of input data to make predictions. ML algorithms learn and optimize their operations. The accuracy of these algorithms increases with the input of fresh data. Different supervised algorithms are used for disease prediction. With the help of a labelled training dataset, the supervised machine learning algorithms are first trained. To categorize the unlabeled dataset into similar groups, the training algorithm is applied to the dataset. Supervised learning algorithms suit well with classification and regression problems.

Support Vector Machines (SVM) is a supervised ML based algorithm. This algorithm can detect and separate features into different classes by using a decision surface. In order to gain insight of the usage of SVM in the healthcare field, [3] we analysed a research document based on the prediction of Chronic Kidney Disease by Almansor et al. where the findings concluded that when comparing two models based on ANN and SVM in predicting, the Artificial Neural Network model performance had higher accuracy. As compared to other machine learning models, SVM tends to produce higher accuracy rates. However, Support Vector Machine has one major drawback; though the accuracy rate is high, it takes ample amount of time for training.

Naive Bayes Algorithm, a prominent ML algorithm is based on Bayes theorem which makes use of a probabilistic classifier. Probabilistic reasoning is used when the information handled by the system is large and complex. [4] A research on Bayes Algorithm was conducted by Pattekari & Parveen. Based on varied input attributes, prediction of heart disease was performed using ML and data mining techniques. As a conclusion of this research, the proposed approach was proved to be the most reliable solution to predict patients ailing with heart disease. However, it has no application in any real-world datasets.

Decision tree is a well-known machine learning algorithm. A decision tree model classifies data items into a tree-like structure. There are multiple levels and in each level there are multiple nodes. The topmost node is known as the root node and the internal nodes represent input values of the tree. We have implemented Decision tree in our disease prediction chatbot, however we got its accuracy less than the KNN classifier.

K-Nearest Neighbors (KNN) algorithm is used for solving classification problems.[3] Based on measures of similarity, it classifies new data points. The KNN algorithm selects a certain distance function and then calculates the distance between the testing and training sample in the multidimensional space, selecting the nearest distance from it. According to the category of the training sample points, the K training sample points are finally judged. KNN algorithm approach is a passive classification method. KNN algorithm trains the sample data by itself, consuming more time in the classification phase thereby taking extremely less time in the training phase. KNN is easy to implement and it possesses classification accuracy significantly high. KNN is dependent on surrounding neighboring samples. For the sample set to be cross-over or overlapped, the KNN method is more suitable than other methods.

Due to the major advantages of the KNN algorithm, a number of researches have been conducted to solve disease prediction problems using the KNN algorithm. We studied two kinds of research based on the KNN algorithm.

In 2016, [5] a heart disease prediction system was proposed by Princy & Thomas using techniques of data mining that focused on the KNN and ID3 algorithms. The results produced by KNN were much better in consistency as compared to other ML algorithms. However, the dataset which was used is unknown. An accuracy of 80.6% was recorded for prediction. Tayeb et al. researched the prediction of heart diseases and chronic kidney diseases using KNN algorithm. In KNN, the most important step is to select the value of K. This is because it can have a substantial impact on the accuracy rate of algorithm.

The research concluded that this method gives a 90% accuracy rate in predicting the diseases; hence better than the earlier studies using the equivalent dataset. However, there are a few common weaknesses of this algorithm as compared to other classification algorithms. The second research proposed that as the algorithm computes the distance of each test instance to all training samples; it will be expensive to use KNN in disease prediction. Another disadvantage was noticed when datasets have low propinquity between various classes, the accuracy rate is primarily affected.

B. Natural Language Processing

The use of NLP has made it easier to understand what exactly a patient has to convey. The major difficulty arises when the system couldn't relate the words of the user and thus unable to produce output. Natural language processing [6] enhances the interactivity among computers and human languages. Understanding of the natural language, natural language generation, and speech recognition are all part of Natural Language Processing. Analyzing the existing

solutions, a web application is developed to predict 24 diseases on the basis of the symptoms provided. The system uses NLP, Decision tree and KNN Algorithms to predict diseases by extracting the input given by the patient. The project has four vital components:

1. Data prediction along with language processing
2. Pathology Report Analysis using Optical Character Recognition
3. Medical Consultation by Doctors
4. Doctor Appointment

3. METHODOLOGY

A. Disease Prediction Chatbot

The chatbot in our project is used for information acquisition. It acquires the patient's information along with the symptoms and the disease is predicted on the basis of the symptoms. The disease prediction chatbot is designed using the concepts of NLP and machine learning algorithms. The chatbot comprises two concepts of Natural Language Processing namely tokenization and wordnet. Upon receiving the symptoms from the user, tokenization is performed and the symptoms are extracted from the sentences the user has entered. Synset is a simple interface that is present in NLTK (Natural Language Toolkit) to look up the words present in WordNet. Synset instances are used to extract words synonymous with the symptoms.

If the symptom entered is incorrect, an invalid response is generated and sent to the user. When the symptoms entered are valid, the extracted symptoms are forwarded to the classifier. The dataset used in the project comprises of a total of 35 symptoms and 24 diseases. Multiple common diseases can be predicted through the classifier. The text exchanged between the user and the chatbot is stored in the database, as it consists of the patient's symptom information. Figure 1 depicts the flow of the chatbot.

The second phase comprises of classification. It consists of two processes, which are learning and prediction. According to the first process, a model is constructed based on the training data. In the second process, the best model is used to predict the response for the data. Our model applies the approach of Supervised Learning.

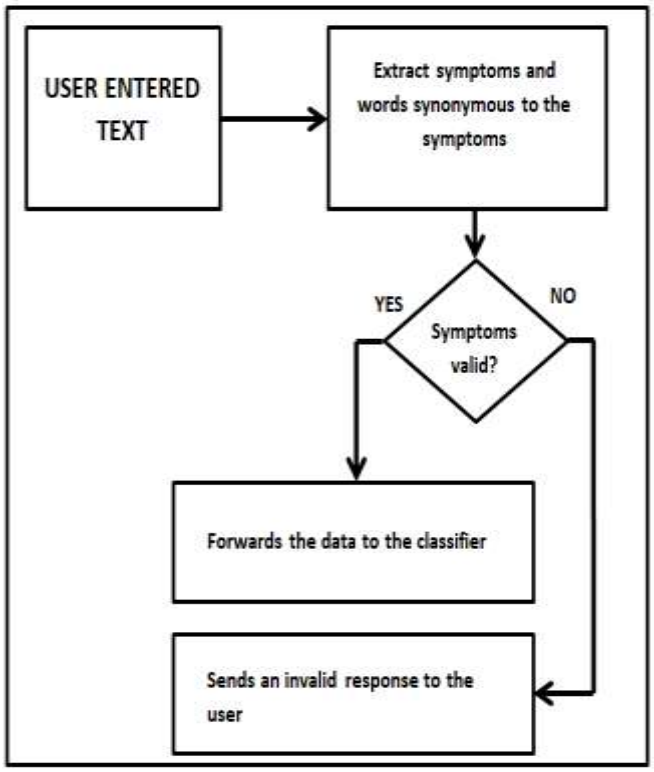

Fig -1: Flowchart of Chatbot

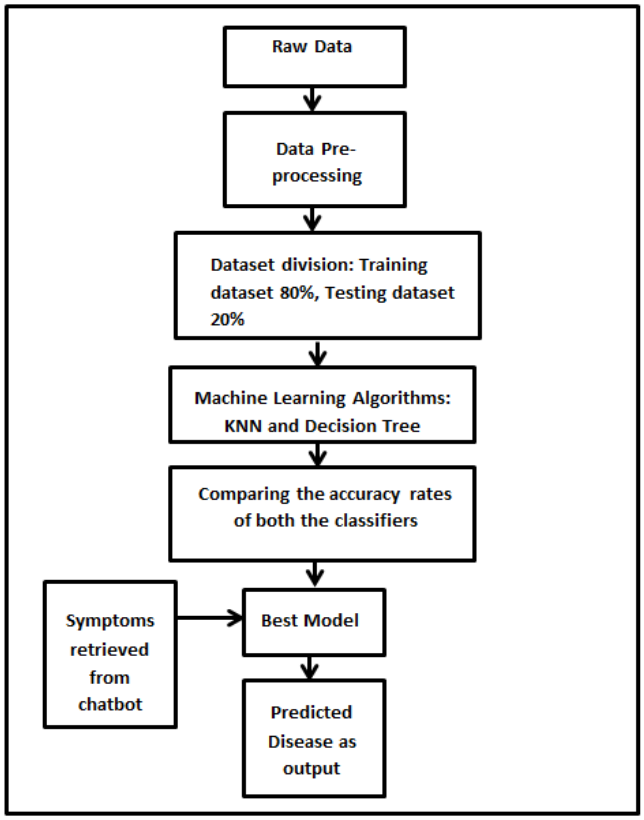

Fig -2: Flowchart of Disease Prediction System

The partitioning of the dataset becomes crucial for getting good accuracy in models. In the model, we have used two algorithms namely KNN and Decision Tree on the training dataset and based on the model's confidence and after testing dataset accuracy, we select the best model algorithm and apply it on the testing dataset to generate accurate results. As per our results, the accuracy of Decision Tree and KNN is 92.6% and 95.74% respectively. Once the classification model is selected, it is applied to the symptoms extracted from the chatbot to predict the disease of the patient.

B. Pathology Report Analyzer

Initially, the analyzer uses the concept of OCR (Optical Character Recognition) thereby converting the scanned image of the report to text. [7] Py-tesseract is an OCR tool by python. The tool recognizes the textual data present in images. Py-tesseract is a wrapper for Tesseract Engine. It is extremely convenient as it processes all the types of image extension supported by the Pillow library. It retrieves the recognized text from the image which can be stored in a variable or it can be written into a file. Tesseract comprises two concepts namely line finding and word recognition. Word recognition is a process in OCR engine where its function is to detect a word and segment it to characters. The rest of the recognition process is applied to non-fixed pitch text. The process of image processing consists of the following stages [8]:

1. Image pre-processing is performed with adaptive thresholding where a binary image is produced.
2. Then the connected component analysis is performed which produces character outlines.
3. The techniques for character chopping and character association are performed to organize the outlines into words.
4. In the end, two-pass word recognition is completed by using methods of clustering and classification.

A patient's pathology report is graphically analyzed by extracting the test results, hence giving the patient a comparative overview of their last reports. Each tested attribute value of the result can be analyzed. The application provides information about the report in an understandable manner and notifies whether the patient's report values lie in the desired range or not. The system gives an alert to the user if any of the value crosses its threshold value.

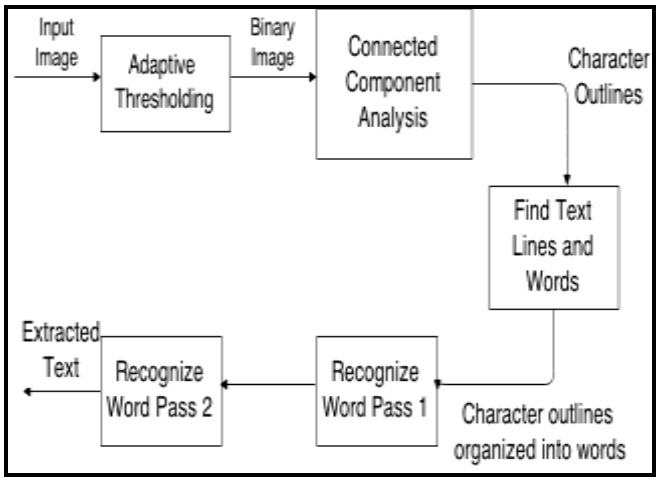

Fig -3: Tesseract Architecture [8]

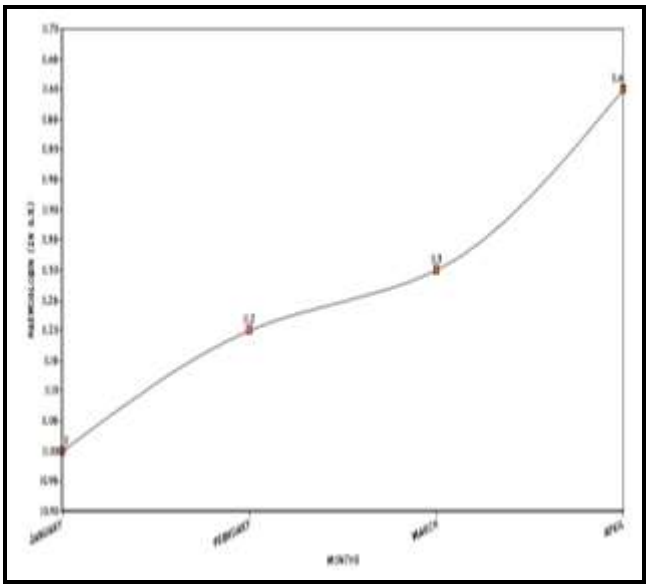

Chart -1: Analysis of patient's haemoglobin value extracted from the pathology report

Additional features present in our web application are:

- Doctor Appointment: The patient can book an appointment based on the specialization of the doctor and the availability of appointment slots.
- Medical Consultation: Doctors can review the medical reports, past prescriptions, symptoms and medical history of the patient and provide consultation accordingly.

4. BLOCK DIAGRAM OF THE SYSTEM

The figures 4 and 5 refer to the block diagram of the proposed system. Figure 4 represents the disease prediction chatbot. Initially, the dataset is classified by two supervised classification algorithms namely KNN and Decision tree.

Based on the accuracy rates, the most accurate classifier is selected as the best model. When the user enters his/her information along with the symptoms in a sentence format, it is tokenized into a set of words with the help of tokenization which is a concept of Natural Language Processing. The tokenized sentence is reviewed to check if it consists of symptoms available in the dataset. If the user enters a word which is synonymous with the symptoms present in the dataset, it is accepted. Synonymous words are retrieved through WordNet. The detected symptoms are forwarded to the best classification model and based on the symptoms; the patient's disease is predicted. The predicted disease is given as an output in the chatbot.

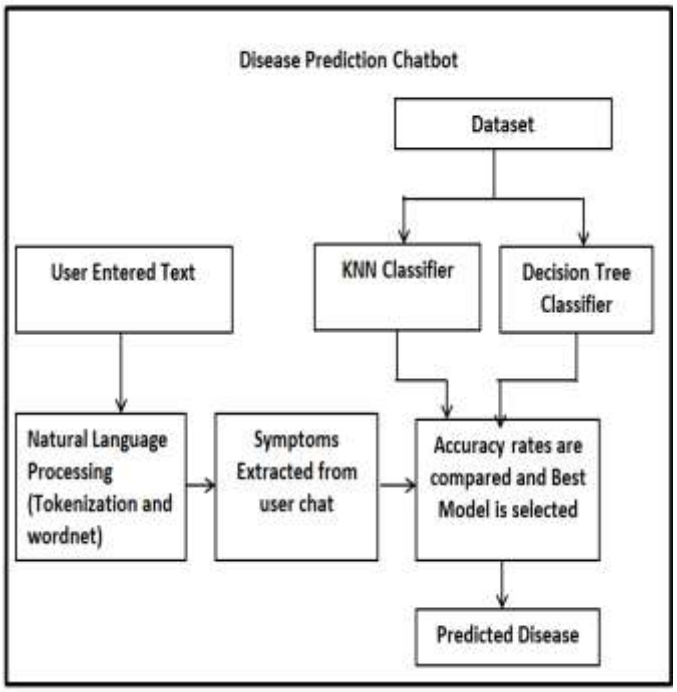

Fig -4: Block diagram of Disease Prediction System

Fig 5 represents the block diagram of the pathology report analyzer. The patient's information is stored in the Firebase Realtime database. The patient can upload their pathology report images in the portal which is stored in Firebase Storage. Once it is stored in the firebase storage, the user can select one or multiple reports for scanning. The image of the report is converted to text using the Optical Character Recognition tool-Tesseract. The extracted text of each report is stored in the database. The extracted text serves two purposes:

1. The user gets notified about their report results. Each result value of the report is analyzed and the system informs result values to the user by translating the report in an understandable manner and informing the user whether the result values lie below, within or beyond its value range.

2. User can select multiple reports to track their test result values. The graphical analysis of each result value retrieved from multiple reports of the user helps in giving an insight of their health and also gives them a comparative analysis. Thus, allowing them to track their improvement from their past test results.

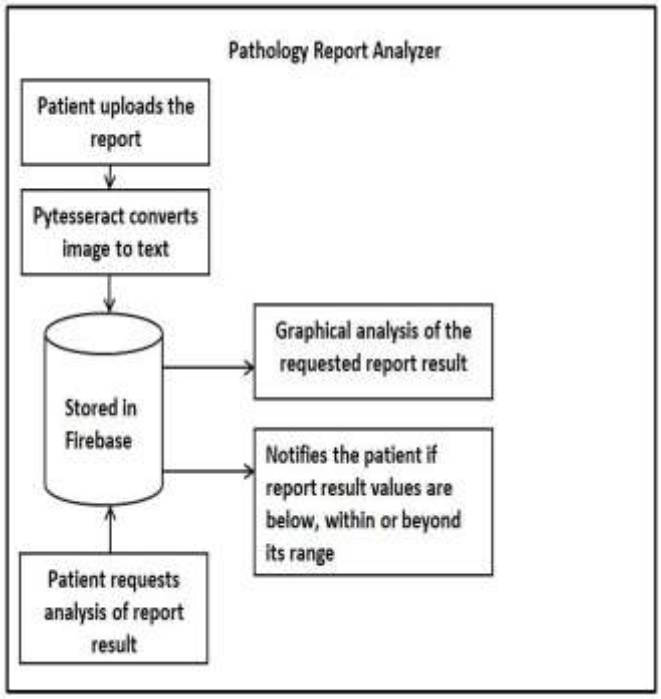

Fig -5: Block diagram of Pathology Report Analyzer

5. CONCLUSION

This paper represents a prediction model driven system that predicts accurate diseases based on the symptoms. The concept of Natural Language Processing is used to design an interactive chatbot to retrieve symptoms provided by the user. The prediction model is designed using Machine learning algorithms such as KNN and Decision Tree. Both the algorithms were applied on the same dataset and based on the confidence and accuracy rate, the best model was selected. As per our results, the accuracy of Decision Tree and KNN are 92.6% and 95.74% respectively. Hence, KNN gave better results in this dataset. The model is predominantly used to classify 24 common diseases. The concept of Tesseract, an optical character recognition tool was used to translate the pathology report images. We integrated classification and OCR techniques in the system in order to provide an application which the patients can utilize effortlessly. We believe that this approach incorporated into existing strategies in the field of healthcare will provide assistance to the health specialists and patients.

6. REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017.

[2] Lalwani, Tarun & Bhalotia, Shashank & Pal, Ashish & Bisen, Shreya & Rathod, Vasundhara. (2018). Implementation of the ChatBot System using AI and NLP.

[3] Yu, Hong. (2019). Experimental Disease Prediction Research on Combining Natural Language Processing and Machine Learning.145-1500.1109/ICCSNT47585.2019.8962507.

[4] Akhil, Jabbar & Samreen, Shirina. (2016). Heart disease prediction system based on hidden naïve Bayes classifier. 10.1109/CIMCA.2016.8053261.

[5] Enriko, I Ketut & Suryanegara, Muhammad & Gunawan, Dinda. (2016). Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters. 8. 59-65

[6] K. Jwala, G.N.V.G Sirisha, G.V. Padma Raju(2019).Developing a Chatbot using Machine Learning

[7] Nair, Akhil., Overview of Tesseract OCR engine(2016).

[8] Ch, Sravan & Mahna, Shivanku & Kashyap, Nirbhay. (2015). Optical Character Recognition on Handheld Devices. International Journal of Computer Applications. 115. 10-13. 10.5120/20281- 2833.

[9] S, Vinitha and S, Sweetlin and H, Vinusha and S, Sajini, Disease Prediction Using Machine Learning Over Big Data (February 2018). Computer Science & Engineering: An International Journal (CSEIJ), Vol.8, No.1, February 2018.

[10] Sarthak Khurana, Atishay Jain, Shikhar Kataria, Kunal Bhasin, Sunny Arora , Disease Prediction System, International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 05,May 2019

[11] Uddin, S., Khan, A., Hossain, M. et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19, 281 (2019).

[12] Patel, Chirag & Patel, Atul & Patel, Dharmendra. (2012). Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications. 55. 50- 56. 10.5120/8794-2784.