AI Based Tumour Detection Using Image and Text Data

Jagreet Kaur[1], Dr. Kulwinder Singh[2]
[1]Research Scholar, IKGPTU, Jalandhar
[2] Professor and Head (IT), GNDEC, Ludhiana

Abstract: In health care system, to detection brain tumour is a challenging task. As it is very difficult to understand images for detection but with the recent advancements in the technologies, it became easy to do the analysis of images and available tools helps to predict the future problems also. Machine Learning and Deep Learning is a subcategory of Artificial intelligence and has the goal to develop various solutions to implement automatic methods to make our systems capable of evolving by themselves. The historical and real time clinical text data and image data are analyzed to determine rules that can be integrated to align software applications. This objective of this paper is to develop an analytical Health care platform for the medical practitioners to administer the patients in the hospital.

Keywords: Healthcare coverage, Deep Learning, data mining, feature selection classification and prediction, medical imaging.

1 INTRODUCTION

Chronic condition is a health disease that is persistent and whose effects are long lasting and adverse. It has major effect on the quality of life of the individual who is affected with it. With the development of intelligent devices for long-term monitoring has the potential to change the delivery of healthcare [1]. Patients may no longer need to make unnecessary hospital visits, previously unknown medical problems might be diagnosed with precision at early stage and patients could possibly be provided with advanced warnings of serious conditions enabling drug delivery to avoid medical emergencies [2]. This technology will be driven by data in which Machine Learning (ML) and deep learning plays an integral role. This paper proposes the use of Machine Learning and Deep Learning to do the real time, predictive and prescriptive analytics for a tumor detection.

The aim of this paper is to create an Artificial Intelligence based healthcare analytical platform using machine learning, deep learning and pattern recognition for analyzing the symptoms and diseases of the patients to provide knowledgeable data to predict and prescribe remedies for the wellness of the patients [3].

1.1 General Context

Nowadays, Health-care costs is very high.
The terabytes or even petabytes of health data available in EHRs present new opportunities and challenges for practitioners that aims to use these data effectively to discover new knowledge to improve health-care [4] [5].Integration of multiple datasets can be broadly classified into two groups: 1) Feature selection and 2) Predictive models.**Feature Selection** is a preprocessing technique that is used to identify the significant attributes, which play a leading role in the task of classification. This leads to the dimensionality reduction.
Predictive models are used to build classification models with high accuracy. Predictive analytics uses technology and statistical methods to search through massive amounts of information, analyze it to predict outcomes for individual patients. Prediction modelling uses techniques such as artificial intelligence to create a prediction profile (algorithm) from historical data.

1.2 Electronic Health Record data

Effectively integrating and efficiently analyzing various forms of healthcare data over a period of time can answer many of the impending healthcare problems [6].

- Clinical notes: Clinical notes contains rich and diverse source of information but it is very difficult to handle clinical notes due to short phrases and ungrammary, abbreviations, misspellings. All information is in semi-structured format.
- Medical Imaging data: The main challenge with image data is that it is not only huge but high dimensional and complex also.
- Medication: Medication data vary in EHR system can be in both structured and unstructured format. Standard codes like National drug code is a unique identifier assigned to each drug by Food and Drug Administration but it is not universally used by EHR system[7]
- Lab results: The standard code for lab is Logical Observation Identifiers Names and Codes(LOINC). Many lab systems still use local dictionaries to encode labs.
- Behavioral data: Capturing the patient's behavioral data through several sensors; their various social interactions and communications. Sensing devices can provide several types of data in real-time.

1.3 Problem Description

Evidence-based medicine is a best treatment method to minimize variation and unexpected costs[8]
Doctors and patients can easily keep track of their health through Electronic Health Record(EHR), a digital record where every individual patient's medical history, demographics, diagnostics tests, etc. are stored and maintained. Suppose a patient has the flu, he just need to go online and enter you're symptoms in the symptom calculator and the software's algorithm will match patients symptoms with others who has had similar symptoms and also show the diagnosis that was most common[9]

1.4 Goal

There are large collection of data related to health are available. Each of these datasets has some patterns that can reveal cause of certain disease. In order to make human life easier, there is a need to extract information from these huge volume of datasets in digital world.[10] Also in every hospital, practitioners use different standards and protocols to generate and store the patient data, so interoperability is also a big issue.The goal of this paper is to develop a interoperable analytical platform that is trained to learn patterns in the data and utilize these patterns to make prediction and prescribe most feasible solutions.[11]

2. OBJECTIVES AND APPROACH

The main Objective and Scope of this paper is to focus to make an AI based analytical platform for tumour detection in health care.This is comparatively a new area in the medical field and therefore research aims to find new methods that could help using of this technology more effectively. Following are the various objectives of this paper are:-

1. To develop data exchange and interoperability architecture to provide a personalized care to the patient.
2. To develop the AI based Analytical platform for integrating multi sourced data.
3. To propose a Predictive and Prescriptive Modelling Platform for physicians to reduce the semantic gap for accurate diagnosis.
4. To compare the proposed system with other state of art techniques.

2.1 General Approach

1. Once a new symptoms enters in the system, the back-end service of the system collects all relevant features of the problem[12].
2. The back-end service than sends these features to the machine learning and deep learning models.
3. The model compares, predicts results and prescribe possible solution.
4. The back-end service receives the predictions and scores, and saves them to data store.
5. Once an practitioners opens a given link, the front-end service triggers the back-end service will retrieve the saved predictions and prescriptions and go through steps 2-4 again.
6. The back-end service returns the list of solutions ranked by the predicted score to the frontend[13]
7. The top three ranked solutions are suggested as a result.

3 METHODS

3.1 Methodology

The overall framework, as shown in fig 3.1, consists of the following four computational steps.

1. In data preprocessing i.e. information extraction, we clean the data, remove missing data, and prepare the data according to the requirement of the task [14][15].
2. In feature selection, we use the statistical measures.[16]
3. In classifier building, we build training and testing sets based on selected features.
4. During evaluation, we characterize the classification and patient similarity schema and determine the results[17].

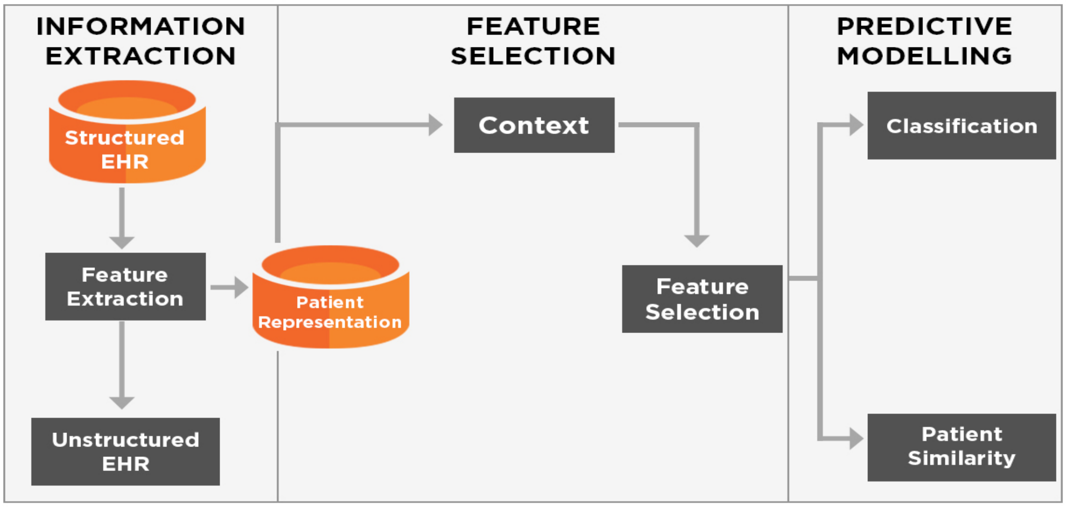

Fig 3.1: Framework

3.2 WorkFlow processes

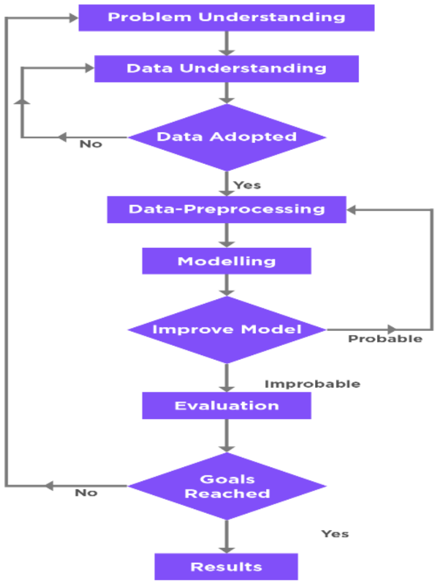

Fig 3.2: Workflow process

It is very important to follow a correct process, as shown in Fig 4.2 from the starting point until the creation and the evaluation of the solution.

1. Problem understanding: Need to consider the problem and way the solution should look like.
2. Data understanding: Take in consideration the amount, the availability, the quality and the quantity of the data. If it is relevant to our current problem.
3. Data preprocessing: Clean the data to focus on what is really important, find the best way for modelling and increase the quality of the data.
4. Modelling: Finding the best method and model that suits to our current problem.
5. Evaluation: Answering the question, if our model fulfil our requirements.
6. Results: Predicted results are calculated [18].

4: DESIGN

This part will describe and explain the conceptual architecture of the project from the machine learning part point of view, data mining, the workflow of the processes for intelligent data analyzing, the choice of the machine learning and deep learning algorithm[19]. We will also justify the choice of the different technologies.

4.1 Conceptual Architecture

The following figure 4.1 shows the conceptual architecture of the proposed analytical platform.

The general idea is that in the frontend of our application, every time a user is using the program and entering the symptoms of the problem, symptoms are transferring into the backend and collecting into the Machine Learning Database [20]. This Database contains the medical history of the previous patients that will be the future input of the solution for our algorithms. The patients symptoms and medical imaging data are compared with the data stored in the database and then predict the future of the patient according to the result generated by the system after the comparison with knowledgeable datasets [21].

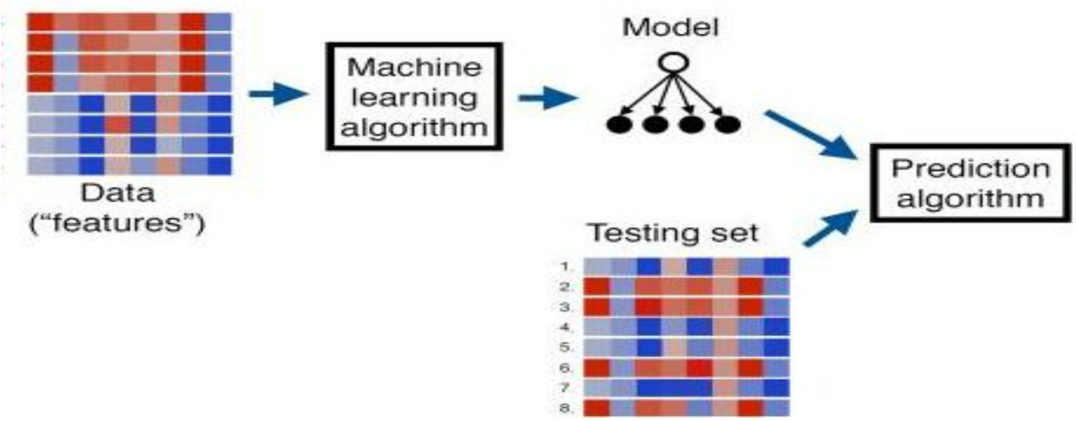

Figure 4.1: Architecture

4. TECHNOLOGIES

This section will describe the main technologies that we will use for this thesis. It is separated in three short parts, the input that contains the data, the machine learning and deep learning and the output.

Input data In the first iteration, we will start by taking a historical text and imaging data in a simple Excel or CSV file and jpeg to save time and concentrate our efforts to establish a proof of concept.

Building the backend with NLP, ML and DL Our machine learning and deep learning model use features extracted from historical patient data, real time data.Since the historical clinical data are useful for understanding about the diseases, we built a NLP pipeline to transform text into useful features for our machine learning models[22]. Similarly, to understand the imaging data, there is a need to understand and extract the features of the images.

Output data Concerning the output of our solution, we will try to generate a predicted functions for the practitioners to retrieve the related information.

5: IMPLEMENTATION

The two first iterations will be performed in local. At the opposite, the third iteration concerns the integration with the other parts of the task.

5.1 1st iteration

The first iteration will focus on the pre-processing phases.

Ingestion of the data: Data ingestion is the process of fetching data from any defined source or device. Data Ingestion can be streamed as real-time or ingested in batches. For building a real-time analytics solution we need to ingest the data in real-time stream.

The Messaging Queue: Handling Real-time data streaming is the most crucial part while building a real-time analytics solution. In fulfill the task, we need to fetch data from several IoT devices. In this case , we will use Apache Kafka for real-time ingestion. Apache Kafka offers high-speed real-time streaming and works on publish/subscribe model. IoT devices can be used to push data to Apache Kafka and again that stream can be used for the further processing.The real-time stream is parallel sent to a database. This data can be used as the historical data for training purposes.

Preprocessing: The pre-processing phase is a crucial phase before the creation of classifier model. It consists of cleaning the different data, removing the noises in the observation and correcting the missing data.

Text preprocessing: As we are working on various documents which may contain different types of information. We will need to do preprocessing very carefully so as not to lose important information. We will follow the steps in terms of pre-processing of data.

1. Part of Speech Tagging: POS tagging basically means tagging every word, a part of speech like noun, verbs, adverb, adjectives. We use POS taggers from nltk package of python like CoreNLPNERTagger, CoreNLPPOSTagger, StanfordNERTagger.
2. Stop Word Filtering: Stop words filtering is very important part of text cleaning as this helps the model to know the import terms in the text like stop words are basically the words which acts as support of objective to form a sentence like is, are, of etc.
3. Stemming: For grammatical reasons, documents are going to use different forms of a word, such as *X-Ray*, *X-Rays*, and *X-ray*. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set.The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
4. Feature Vector Extraction We will be doing feature extraction like as follows:

TF-IDF and n-gram: This is the frequency of occurrence of a word in a text ignoring most frequently occurred words in text. Similarly, word co-occurrence methods manage statistical information about the number of times a word has happened and the number of times it has happened with another word.

5. Similarity Matrix: It helps in finding similarity between words like medicated or vaccinated etc.
6. Positional Feature :Positional value of a sentence is also extracted. A sentence is relevant or not can also be judged by its position in the text.

Image preprocessing

1. Annotation of Image Data: Image data is obtained in the form of jpg format. Therefore, annotation of image data is required. A python library is used known as labellmg or LabelMe matlab tool for annotating the required object from images. While annotating, the image is obtained in the form of xml files.
2. Fetching of Data from xml files: In this stage, the data is present in xml format. Therefore, for each annotated object the values of coordinates are fetched.
3. Convert the shape of Annotated object into Bounding box: After fetching the values of coordinates, the shape of annotated object has to be verified and convert the obtained shape into bounding box. This is because, Tensorflow Object detection API supports annotated objects in the form of bounding box only.
4. Fetch the size of jpg images: After conversion, there may be the possibility that there is no information about the size. Therefore, fetch the image size using pillow python library and merge it with the obtained current data.
5. Convert the data into .Record file format: After fetching the complete required data, the data is converted into the binary encoded file because model and API works with this required format.
6. Image Augmentation: In the case of healthcare, the number of images may not be sufficient. Therefore, data augmentation is implemented to increase the number of images by performing various operations such as rotation, shearing, cropping, mirroring, etc.

5.2 2nd iteration

The second iteration contains the creation of the model, the training phase and the evaluation of our results.

Text Analytical Model

We will be building a sequence to sequence model for text summarization referred by google team as Automatic Text Summarization.

Image Analytical Model

Various number of models are available for object detection but two models are selected due to their better accuracy. They are:

1. Faster RCNN: this model is suitable for the detection of objects with good accuracy. This is the one of the convolutional neural network along with the feature of region based training.
2. Generative Adversarial Network: this model is also suitable for object detection. But, some region based networks are not able to detect small objects with good accuracy. Therefore, for the detection of small objects along with region based feature this model is recommended.

Training the data on the model

To train our model, we will partition our data into two datasets, one for the training, and the other for testing our classifier model. We will configure it to have a partition with 80% of the amount of the data for the training and 20% for the prediction test.

Evaluation of the model

To evaluate the score of our model, we will use the Precision and Recall metrics.

7.3 3rd iteration: To identify the best possible recommendations for problem, we apply a learning-to-rank approach and build a retrieval-based pointwise ranking algorithm.

8 RESULTS

The experiment will be carried out on health care text and cancer dataset. The outcome of the experiment is evaluated using the parameters Precision, Sensitivity, Accuracy and specificity and it is compared with the existing technique called Genetic algorithm (GA).

- Specificity –measures the proportion of negatives that are correctly identified.
- Sensitivity- measures the proportion of positives that are correctly identified
- Accuracy – Determines the correctness
- Precision –Repeated process same result

CONCLUSION

The objective of this paper was to make the healthcare system interoperable. Further, the system can predict the disease and prescribe the most appropriate medication. The goal was to develop an analytical platform to provide the suggested information to frontend applications. We determine the best way and technologies to establish a proof on concept that shows by analyzing the historical text and medical imaging data and used machine learning and deep learning algorithms to make a system AI based.

ACKNOWLEDGEMENT

Authors are highly thankful to the department of RIC, IKG Punjab Technical University, Kapurthala, Punjab, India for providing the opportunity to conduct this research work.

REFERENCES

1. M. Roopa, Dr.S. Manju Priya, “A Review of Big Data Analytics in Healthcare” in Proc. Conf. International Journal for Scientific Research & Development, Sp. Issue – Data Mining, 2015.
2. Rotsnarani Sethy, Mrutyunjaya Panda, “Big Data Analysis using Hadoop: A Survey” in Proc. Conf. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, July 2015.
3. Dalia AbdulHadi AbdulAmeer, “Medical Data Mining: Health Care Knowledge Discovery Framework Based on Clinical Big Data Analysis” in Proc. Conf. International Journal of Scientific and Research Publications, Volume 5, Issue 7, July 2015.

4. Telmo da Silva Morais, “Survey on Frameworks for Distributed Computing: Hadoop, Spark and Storm” in Proc. Conf. Doctoral Symposium in Informatics Engineering, 2015.
5. Abdur Rahim, Mohammad, “Big Data for Context-aware Monitoring – A personalized Knowledge Discovery Framework for Assisted Healthcare”, in proc. Conf. Transactions on Cloud Computing, IEEE, 2015.
6. D. Blum, S.X. Raj, R. Oberholzer, I. I. Riphagen, F. Strasser, S. Kaasa ,"Computer-based clinical decision support systems and patient-reported outcomes: a systematic review",in Proc. Conf. The Patient-Patient-Centered Outcomes Research, vol. 8, pp. 397-409, 2015.
7. Akshay Raul, Atharva Patil, “Knowledge Discovery, Analysis and Prediction in Healthcare using Data Mining and Analytics”, in Proc. Conference, 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.
8. Aditi Bansal and Priyanka Ghare, “Healthcare Data Analysis using Dynamic Slot Allocation in Hadoop” in Proc. Conf. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-3 Issue-5, November 2014.
9. Van-Dai Ta, Chuan-Ming Liu, Goodwill Wandile Nkabinde,” Big Data Stream Computing in Healthcare Real-Time Analytics”, in Proc. Conf. International Conference on Cloud Computing and Big Data Analysis, IEEE, 2016.
10. G Rajesh Chandra “Tumor detection in brain using genetic algorithm”, in Proc. Conf. 7th International Conference on Communication, Computing and Virtualization ,2016.
11. N. Senthilkumaran, R. Rajesh, “A Study on Edge Detection Methods for Image Segmentation”, in Proc. Conf. Computer Science (ICMCS, Vol. I, pp.255-259), 2009.
12. Malathi R,Dr. Nadirabanu Kamal A R, “Brain Tumor And Identification Using K-means Clustering Technique,” proceedings of the UGC sponsored national conference on advanced networking and applications, 27th march, 2015.
13. Mohammed ElmogY, Rashid Al-Awadi, Eman Abdel-Maksoud, ”Brain tumor segmentation based on a hybrid clustering technique” , in Proc. Egyptian Informatics Journal, 2015.
14. N. Sharma, A. Ray, S. Sharma, K. Shukla, S. Pradhan, and L. Aggarwal, “Segmentation and classification of medical images using texture-primitive features: application of BAM-type artificial neural network,” in Proc. Journal of Medical Physics, vol. 33, no. 3, pp. 119–126, 2008.
15. A. Chaddad, “Automated feature extraction in brain tumor by magnetic resonance imaging using gaussian mixture models,” in Proc. International Journal of Biomedical Imaging, 2015.
16. Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang, “Deep Learning for Health Informatics” in Proc. journal of biomedical and health informatics,IEEE, vol. 21, january 2017.
17. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N., “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”, in Proc. International Conference on Knowledge Discovery and Data Mining, ACM, August 10-13, pp. 1721–1730, 2015.
18. Akshay Raul, Atharva Patil, “Knowledge Discovery, Analysis and Prediction in Healthcare using Data Mining and Analytics”, in Proc. Conference, 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.
19. K. Bahwaireth, L. Tawalbeh, "Cooperative models in cloud and mobile cloud computing", in Proc. 23rd International. Conf. in Telecommunication, pp. 1-4, 2016.
20. Witten me, Frank E, Hall M, Pal C. , “Data Mining: Practical Machine Learning Tools and Techniques”, 4th edition. Burlington, MA, 2016.
21. Luo G., “A review of automatic selection methods for machine learning algorithms and hyper-parameter values.”, New Model Analytics Health Inform Bioinformatics, 2016.
22. Herland, Matthew, Taghi M. Khoshgoftaar, and Randall Wald., "A review of data mining using big data in health informatics." in Proc. Journal of Big Data, 2014.