Disease Prediction using Machine learning techniques

# Abstract

This Dissertation traverse the impact of Machine Learning through an interactive website. With the growth of big data in healthcare communities, analysis of medical data benefits early disease detection and patient care. The Accuracy of the analysis reduces when there is incomplete medical data. The healthcare sector, like other sectors is driven by data and this data are gathered in real time, observed, stored and evaluated by machine learning techniques to offer better detection of diseases. There is a need of proper medium through which patients could get the information earlier from the symptoms of the disease they have. The normal procedures followed by doctors involves traditional differential diagnosis which takes time and less predictive than the disease detection performed by machine learning techniques. The study was conducted on mix of statistical and predictive approaches. In this research work, three machine learning techniques such as Decision Tree, Random Forest and Naïve Bayes is focused on to explore how these techniques help to effectively predict the diseases. The application of three aforementioned technique is shown in the literature review and how it is far better than the traditional approaches of the doctor.

# Chapter 1: INTRODUCTION

## BACKGROUND

This dissertation proposes to predict disease of an individual in Bangalore city located in Karnataka, India using Machine Learning techniques. Disease is particular abnormal condition that negatively affects structure or function of part or all of an organism that is not due to any external injury. Diseases are often elucidated as medical conditions associated with signs and symptoms which are specific to that disease.

## AIM AND OBJECTIVE

The main aim of the research is to the outline the application of machine learning techniques to detect the diseases in humans. Thereby, particular user could be aware and should take good care of himself if there are symptoms of a disease. Machine learning techniques is termed as the application of artificial intelligence that provides the opportunity to the system for the ability to automatically learn and enhance from the experience gathered without being programmed to perform these kind of works (Witten, et al., 2016). The process of machine learning focuses on development of designing computer program that access data in terms of determining useful information from it. This research paper is based upon predicting disease from machine learning technique, in most of medical and health care sector machine learning algorithm are used for accurate disease prediction for the purpose of collecting hospital data and storing it for future usages. Moreover, algorithm in machine learning are also being used for calculating fundamental information in disease prediction which includes detecting life routine and medical conditions.

The main objectives of this research are the following:

* To analyze and predict the diseases in a patient from the symptoms shown in Bangalore, the capital city of Karnataka using different machine learning techniques.
* To develop a website where a particular can easily detect the disease from the symptoms shown in him.

## The search for the causes of human diseases goes lower back to ancient past. Hippocrates, a Greek medical doctor of the 4th and 5th centuries BCE, is assigned with being the first to adopt the concept that disease isn't always a visitation of the gods however it is caused by earthly affects. Scientists continually searched for the causes of disease and, indeed, have observed the causes of many. With few exceptions Diseases do not appear overnight, In fact Disease develop slowly. The main causes of disease are Injury like cellular damage by unhealthy foods, Stress and Emotional injury. Toxicity like Exogenous (chemicals outside the body) and Endogenous (chemicals inside the body).Also deficiency in Nutrients, Sunlight, and Air etc. are some of the causes for diseases. These causes develops a disease and can be taken care before affecting with a disease by introducing a proper medium like website which helps to discover the disease from the symptoms. After unravelling the disease, a particular should take good care of himself in the primary stage before getting affected to a disease fully. The primary research questions that this research work would focus is addressed below:

* How accurately can the diseases be detected through machine learning techniques and can be implemented on a website?
* Examine the prediction accuracy in different sectors on end to end user demands?

# ORGANISATION OF DISSERTATION

# Chapter one deals with the background of topic, defining the problem and its causes. This section also discusses about the research question, aim and objectives, scope and limitations of the research. Chapter two discusses about Literature review, which includes the significance for insertion of each topic. Chapter three contains research methodology and methods featuring strength and weakness of the topic of dissertation. Chapter four deals with artefact design and development, showing how artefact contributes to the dissertation. Chapter five illustrates the findings and test results of each algorithm. Chapter six scrutinize the discussion part, where the work is reviewed. This section also deals with Interpretation of results, answering the research questions and discussing the conclusion of the findings. Chapter seven deals with general conclusion by summarizing the discovery which also includes future work recommendations. Chapter eight and nine enumerates the bibliography and appendices respectively.

* 1. DISSERATION SCOPE AND LIMITATIONS

The scope of this dissertation is to predict the diseases in an individual by the symptoms shown in them in the city of Bangalore, Karnataka. In this dissertation, a website is created where user can predict the disease based on the symptoms shown by the patient. The website

# Chapter 2: Literature review

## The concept of machine learning techniques and its application in different sector

According to (Witten, et al., 2016), machine learning is described as the artificial intelligence application which allows the systems to automatically understand, learn and enhance from the experience without being programmed to do such things. It allows the system to process data and learn from the information gathered from observation. Machine learning mitigates human intervention and thus adjust accordingly to the situations (Huang, et al., 2015). It is also often illustrated as the study of statistical models and algorithms that are being used by the computer to perform certain tasks effectively. Machine learning algorithms generally develop mathematical models based on sample data to make sure that accurate predictions can be made without being programmed to do it.

From the article by (Lee, 2013), machine learning is categorized into different parts such as supervised learning, unsupervised learning, reinforcement learning and semi-supervised learning. Supervised machine learning considers labelled training data for learning about models and helps in detecting the outcome related to unseen or future data. The supervised learning method involves steps such as data preparation, training step, test step and production development (Oquab, et al., 2015). Certain examples of supervised learning are voice assistants by Apple Siri, Microsoft Cortana, Gmail filters, predictions done by weather apps etc. Unsupervised learning explains only about the input data and not about output variables. The application of unsupervised learning is seen in distributing the data or modelling the underlying structure to have more knowledge related to data. Based on the article by (Criminisi, et al., 2012), semi-supervised machine learning is used where there is a massive amount of input data but only a few would be labelled as data. Reinforcement learning is another kind of machine learning which allows the system to monitor the environment and gain knowledge related to the ideal behavior to maximize the idea of cumulative reward (Jiang, et al., 2016).

Application of machine learning can be seen in different sectors. According to (Patel, et al., 2015), the financial sector uses machine learning for getting insight on data related to investment opportunities and to circumvent fraud. The retail sector uses the machine learning to offer a better shopping experience for the consumer and the algorithms help to differentiate the consumer data and get better understanding related to the requirements of the consumers. Similarly, the automotive sector uses algorithm techniques for enhancing the operations, marketing and a better understanding of the requirements of the consumers.

According to Forbes, (2019), the application of machine learning in the manufacturing sector has brought revolution as because it has improved the performance of semiconductor manufacturing yield by 30%, resulting in optimizing fab operations and reducing scrap rates. Furthermore, it has also been identified that machine learning such as the implementation of AI in the supply chain has reduced the errors by 50%, which has increased the productivity within the management. Manufacturing industries in worldwide are considering the use of machine learning as because it helps them in automation testing which has not only increased the rate of defect detention by 90%, but it has also helped in storing materials and calculating its value to increase productivity. The retail industry also uses machine learning to design their product and services in terms of improving the quality along with offering innovativeness in terms of retaining the attention of the customers. As stated by Prnewswire, (2019) machine learning in retail industry also helps in calculating annual sales turnover and sales pattern through the help of big data analysis, which helps the retail industry in analyzing trends that could help them in growing their business in future. Machine learning technique in the retail industry also helps in improving the transactional process for the customers. AI-enabled technology improves the online transaction process for the customers which provides opportunities of making fast transactions, keeping customers information safe and providing a wide array of information for increasing customer trust towards the retail industry (Business-standard, 2019). Therefore, the application of machine learning in different sectors becomes essential to upheld strong performance and productivity.

## The application of machine learning in the healthcare sector

From the article by (Durairaj & Ranjani, 2013), in the data-driven healthcare sector, machine learning has become the top trend because of the use of sensors and wearable devices which can store and access data of the patients in real-time. Vital signs of the patient's health can be monitored with the help of these sensors and then this information is being processed in machine learning analytics hub to outline anomalies and offer a better approach to the doctors for treatments (Zhang, et al., 2015). It also helps to provide better diagnoses of the patients by monitoring the trends gathered during a period (Elhoseny, et al., 2018).

Healthcare is considered as one of the most important industry as because it provides value-based care to millions of people, which helps in improving the lifestyle of people and at the same time contributing towards the development of the country. As stated by (Mckinsey, 2019) it has been identified that the healthcare industry in today's world is developing alternate staffing models, provides smart healthcare services and reduce administratively and supply chain cost. Technology giant Google recently developed a machine learning algorithm which helps in identifying skin cancer. Moreover, machine learning in the healthcare industry also helps in identifying thousands of data points, helps in identifying risk, and helps in resource allocation and provide outcomes that help in improving the infrastructure of the healthcare industry (Natarajan, et al., 2017). One of the primary application of machine learning technique is that it helps in detecting early-stage drug discovery process which includes R&D technologies. The R&D process in healthcare industry helps in generating precision medicine along with determining alternative paths for therapy in multifactorial diseases. According to Shafique, et al., (2015) machine learning in the healthcare industry also helps in developing predictive analytics that aids in developing a more personalized medical treatment for the patients. IBM Watson machine learning technique is at the forefront of in terms of leveraging patient medical history in terms of generating multiple treatment options such as sophisticated health measurement capabilities and ML-Based healthcare technologies that work towards providing better medical facilities to the patients (Yang & Lee, 2016). Henceforth, the application of machine learning in healthcare industry helps in improving the diagnoses process along with providing better medical facilities.

## The normal procedures of disease detection

Disease detection or diagnose is defined as the process of evaluating the patients' condition or disease by monitoring the signs and symptoms. The physical examination and the history of the patients are being observed for getting the required information on diagnosis. Due to the nonspecific nature of the symptoms and signs, diagnosis often becomes challenging for the doctors and therefore, it leads to the development of differential diagnosis wherein different possible explanations are contrasted to have a piece of better information related to the symptoms shown by the patients. Differential diagnosis comprises of correlation of different pieces of information which is then trailed by differentiation and identification of patterns, as suggested by (Isola, et al., 2012).

According to Yang, et al., (2015) the normal procedure of disease detection comprises of diagnostic procedures which comprises of physical examination, medical tests and posthumous diagnosis are being considered before diagnosing a patient. In the initial stage of patient diagnosing it becomes essential for the medical facilities to determine the diseases, signs and symptoms of the patient and then perform a medical examination on them. It has also been identified that the normal procedure of disease detention also comprises of challenges such as failure to identify the symptoms and signs due to non-specific nature of the disease (Warner, et al., 2019). The normal procedure of disease detection also involves correlation of various piece of information which are followed by recognition and differentiation of patterns. The procedure of diseases detection includes detecting any kind of deviation such as anatomy which includes the structure of the human body, physiology which includes how the body works and pathology which includes what can go wrong in physiology and anatomy. The normal process of diseases detection also includes measuring the current condition of the patient and assist them in determining the process of diagnostic. According to French, et al., (2018) the diseases detection of the patient also comprises of gathering information about the patient, which includes medical history, diagnostic test and physical examination. Additionally, the process of disease detection also includes diagnostic of medical test which is performed to detect the disease along with providing prognostic information that helps in improving the medical condition of the people (Sakr & Elgammal, 2016). Pattern recognition is also an essential part of diseases detection where physicians recognize a pattern that includes clinical characteristics in terms of identifying signs of certain diseases or conditions.

## How diseases can be detected through machine learning techniques such as Decision Tree, Random Forest and Naïve Bayes.

The tremendous advancement of the big data in recent years has allowed the health sector to make promising development in predicting diseases. Execution of machine learning in the health care sector shows promising results whether in case of detecting multiple diseases or epidemic outbreak. Through the application of machine learning, doctors have been able to predict the diseases effectively and offer better solutions to the patients (Ghoneim, et al., 2018). Machine learning techniques such as Decision Tree, Random Forest and Naïve Bayes have come in handy to detect the diseases on time. According to (Witten, et al., 2016), decision tree machine learning technique is a visual chart that is normally structured in a hierarchical order. Root tree is the tree’s topmost node and on the other hand, a test on the attribute is represented by generation of an internal node done by splitting the algorithm (Mohamed, et al., 2012). This machine learning technique falls under the supervised learning technique and the application of decision tree helps to map out items which in turn provide predictive results. This technique falls under the classification model that segments the input data into a result. This result might then be directed as a particular class domain. Development of the classification model as a decision tree is done through training medical dataset.

According to Maji & Arora, (2019) decision tree as a machine learning technique helps the healthcare industry to take the decision that is used to improve the overall medical procedure. Decision tree in healthcare industry helps in predicting diseases such as it identifies the type of disease that the patient is suffering from and provide guidance to the physicians to take suggestive actions. The implementation of the decision tree as machine learning also allows the physicians to better identify the most favorable options for the patients, furthermore, it also provides access in terms of yielding information that could be used in terms of reducing risk among the patients (Salih & Abraham, 2015). Decision tree in medical health also provides access to graphical and logical sequences that enables the physicians to determine the outcome of the patient’s life year’s expectancy. Once, the decision tree has been mapped it helps the physicians in determining values, outcomes and probabilities through which best healthcare choice are being identified that helps in improving the health status of the individuals. Additionally, it has also been identified that decision tree analysis in healthcare also becomes important because it also develops other utility models which include time trade-off and standard gamble. These two utility models are the ways of assessing how a patient feels about their current state of health as compared to the hypothetical possibility of wellness.

As stated by Swain, (2016) Decision tree analysis in the healthcare industry also helps in performing sensitivity analysis, this analysis is done through systematically varying values that are performed through a credible range. The process of analytics n decision tree process helps in introducing different ranges and structures which are used to measure the probability distribution of patients. Decision tree in healthcare comprises of four classical approaches, one hybrid approach and an evolutionary approach which assist the physicians in terms of determining the disease that the patients are suffering from along with providing them with the treatment that cures them within the limited period. In more specific terms decision tree as machine learning technique in healthcare supports the early and accurate diagnosis of myocardial infarction that helps in detecting the type of disease that the patient is suffering from (Barros, et al., 2015). Decision tree as machine learning within the healthcare industry helps in identifying the particular type of decision problem, which includes diagnostic and intervention type decisions, the decision analysis might be useful for the physicians to make evidence-based decisions. Therefore, the implementation of the decision tree as machine learning in disease prediction helps in diagnosing people and improve their medical health status.

On the other hand, Naïve Bayes is often highlighted as the classification technique which is drawn based on Bayes theorem, as stated by (Bijalwan, et al., 2014). Bayes algorithm is relatively easy to develop and highly useful for exploring data from a huge data set (Chen, et al., 2011). The Naïve Bayes allows the users to change the data set to the frequency table. After this, it also develops the likelihood table by exploring the probabilities such as overcast probability. Naïve Bayes is used in the health care sector to predict accurate results from the provided massive volume of data. It offers methods to calculate the posterior probability. Its classification normally assumes that the impact of a predictor on give n class is not dependent on the other predictors' values and this assumption mentioned above is known as class conditional independence.

According to Singh, et al., (2017) Naïve Bayes as machine learning in disease prediction helps the physicians in maintaining general health records that provide sensitive and important data concerning identifying symptoms and diseases. Naïve Bayes also comprises of training labelled datasets that are used to predict whether the system is providing correct data or not. Moreover, analysing the Naïve Bayes it has also been identified that multi-class learning model also plays a crucial role in terms of classifying binary probabilistic that are used to extract positive datasets and helps in determining the class decision. Algorithms in Naïve Bayes needs to be trained on labelled data sets, which are used for supervised learning and can provide correct classification and prediction of disease symptoms (Alonso, et al., 2017). The Naïve Bayes calculates the prior probability of disease that might occur based on the symptoms. As per the process of disease prediction in Naïve Bayes, it uses pre-processing task which includes tokenization which includes providing correct credentials which help in correctly diagnosing the patient. Moreover, it has also been identified that Naïve Bayes as machine learning in the healthcare industry is it is used to store the updated data by the administrator from time to time. Naïve Bayes technique will also provide flexibility to the physicians to faster access the medical diagnosis in terms of improving the medical health status of the people. Doctors can also login into the system in terms of identifying the particular diseases that can be used to improve the medical conditions of the patient.

As stated by Nahar & Ara, (2018) Naïve Bayes in healthcare industry designs and develop algorithms that are used to determine inference diseases along with its respective medication that helps in determining diseases. Moreover, it has also been identified that Naïve Bayes in medical detection also helps the physicians to access secure and trustworthy platform for better advice on healthy routine diets, medical diagnosis and better interaction with the patients. Naïve Bayes in disease detection also helps the physicians in terms of developing Java swing application that is managed through MySQL database in terms of providing faster access to leverage a large amount of data that are accessible for disease detection (Vembandasamy, et al., 2015). Naïve Bayes algorithm in disease detection also helps in data mining and setting classification of data that becomes useful for detecting diseases among the patient and treat them better.

Moreover, according to (Zhang & Ma, 2012), Random forest uses the cart model bootstrapping algorithm and it develops many trees with multiple initial variables. It normally explores a sample of 100 observations and out of which 5 random samples are generally selected as the initial variable to develop the cart model. This process is being repeated by Random Forest for 10 times and then it comes to the final prediction. Weka tool is normally considered as the machine learning tool where this process related to the Random Forest can be classified, predicted and missing values can be found.

According to Kumar, (2016) random forest as machine learning in disease detection helps the physicians in designing task-specific objective function along with developing adapted posterior model. Moreover, it has also identified that random forest in disease detection also helps in forming computer vision along with medical images that are highly multi-dimensional that provides brief detail images which are used for medical diagnosis process. Random forest technique in disease prediction is also used for advanced medical treatment procedure such as ultrasound and annotation results that provides results which are used for detection of disease among the patient (Jabbar, et al., 2016). Moreover, random forest techniques are also used to supervise the activities within the medical facilities such as keeping the records of the patients along with providing them better medical facilities that help in improving the overall treatment facilities. Furthermore, random forest as machine learning in disease prediction are also useful as because it helps in analyzing patient condition through medical imagery including computer-aided diagnosis, tracking image and segmenting the categories of patients so that their medical conditions can be identified and better medical treatment facilities can be provided to them (Mullainathan & Spiess, 2017). Analyzing the random forest disease it has also been identified that physicians use this machine learning technique as because it helps them in measuring the type of disease along with its related impact on the patients which helps in diagnosing them along with improving their health status over the period. As stated by Belgiu & Drăguţ, (2016) random forest technique also uses computer-aided diagnosis that is undertaken by the medical experts in terms of examining patient data along with taking care of their health condition. The computer-aided diagnosis through random forest helps the physicians in screening the health status of the patients along with identifying the type of disease that the patient is suffering from in terms of providing them with the medical treatment that helps them in overcoming their disease and lead a happy and healthy life.

Chapter 3: Methodology

3.1 Dataset

The dataset used for this project is the Healthcare dataset which was obtained from Bangalore Hospital. The dataset contains about 5000 different patients with symptoms shown by them. In the dataset, there are around 132 unique symptoms and 41 unique diseases. The snapshot of raw data received from the hospital is attached below, where 0,1,2… represents Patient ID and the next columns provide information about the symptoms shown by the patient and final column gives the disease name.

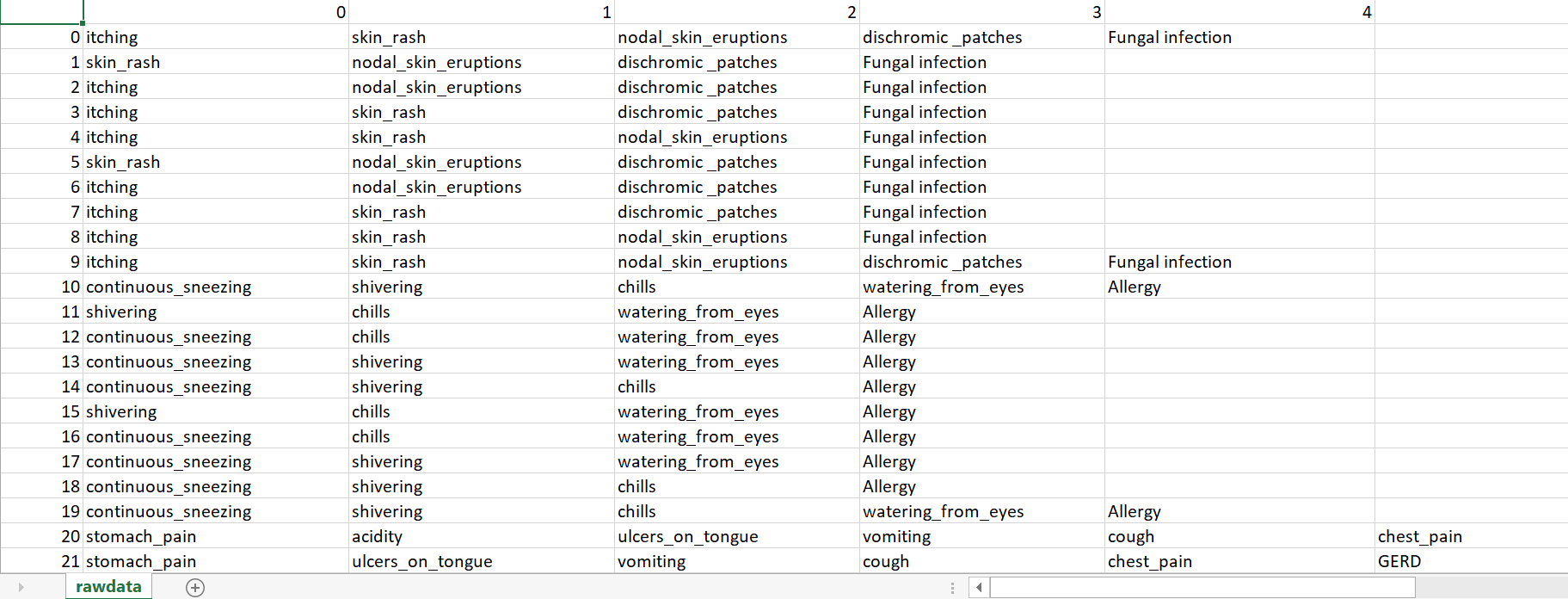


Fig 3.1 Snapshot of the raw dataset received from the Hospital.

3.2 Data Pre-processing.

The tool used in this project to work with the data for pre-processing was Python-3.Python is an interpreted, object-oriented, high level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together(Python: What is Python?, no date). Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed.(Python: What is Python?, no date).

Once the data was available from the hospital the most important or rather the most crucial step before applying any sort of algorithms, was to perform pre-processing on the data so as to convert the information to a machine-readable format.