Disease Prediction using Machine learning techniques

# Abstract

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. 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. 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. In the first chapter of the research study, aim, objectives and questions related to this research has been framed. In the second chapter, these objectives are broadly discussed to gain insight on the concept of machine learning, the techniques and its applications in different sectors, especially health care sector. 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

## Introduction

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

## Research aim

The main aim of the research is to the outline the application of machine learning techniques to detect the diseases. Machine learning techniques such as Decision Tree, Random Forest and Naïve Bayes have been used to explore the disease detection process.

## Research objectives

The research objectives for this study are as follows:

* To understand the concept of machine learning techniques and its application in different sector
* To evaluate the applications of machine learning in health care sector
* To determine the normal procedures of disease detection
* To examine how diseases can be detected through machine learning techniques.

## Research questions

The research questions are as follows:

* Describe the concept of machine learning techniques and its application in different sector.
* What are the applications of machine learning in health care sector?
* What are the normal procedures of disease detection?
* How can the diseases be detected through machine learning techniques?

# 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 the 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 on the basis of 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 labeled 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 massive amount of input data but only few would be labeled 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 the 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 helps to differentiate the consumer data and get better understanding related to the requirements of the consumers. Similarly, automotive sector uses algorithm techniques for enhancing the operations, marketing and better understanding to the requirements of the consumers.

## The application of machine learning in healthcare sector

From the article by (Durairaj & Ranjani, 2013), in the healthcare sector which is data driven, 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 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 of time (Elhoseny, et al., 2018).

## 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 is being clearly observed for getting 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 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).

## 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 the 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 decision tree is done through training medical dataset.

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

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 in order 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.

# References

Bijalwan, V., Kumar, V., Kumari, P. & Pascual, J., 2014. KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application,* 7(1), pp. 61-70.

Chen, C., Zhang, B. & Vachtsevanos, G., 2011. Prediction of machine health condition using neuro-fuzzy and Bayesian algorithms. *IEEE Transactions on instrumentation and Measurement,* 61(2), pp. 297-306.

Criminisi, A., Shotton, J. & Konukoglu, E., 2012. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision,* 7(2-3), pp. 81-227.

Durairaj, M. & Ranjani, V., 2013. Data mining applications in healthcare sector: a study. *International journal of scientific & technology research,* 2(10), pp. 29-35.

Elhoseny, M. et al., 2018. Secure medical data transmission model for IoT-based healthcare systems. *IEEE Access,* 6(1), pp. 20596-20608.

Ghoneim, A., Muhammad, G., Amin, S. & Gupta, B., 2018. Medical image forgery detection for smart healthcare. *IEEE Communications Magazine,* 56(4), pp. 33-37.

Huang, G., Bai, Z., Kasun, L. & Vong, C., 2015. Local receptive fields based extreme learning machine. *IEEE Computational Intelligence Magazine,* 10(2), pp. 18-29.

Isola, R., Carvalho, R. & Tripathy, A., 2012. Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and $ k $-NN. *IEEE transactions on information technology in biomedicine,* 16(6), pp. 1287-1295.

Jiang, C. et al., 2016. Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications,* 24(2), pp. 98-105.

Lee, D., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on Challenges in Representation Learning, ICML,* 3(1), p. 2.

Mohamed, W., Salleh, M. & Omar, A., 2012. A comparative study of reduced error pruning method in decision tree algorithms. *2012 IEEE International conference on control system, computing and engineering,* 1(1), pp. 392-397.

Oquab, M., Bottou, L., Laptev, I. & Sivic, J., 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1(1), pp. 685-694.

Patel, J., Shah, S., Thakkar, P. & Kotecha, K., 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications,* 42(1), pp. 259-268.

Witten, I., Frank, E., Hall, M. & Pal, C., 2016. *Data Mining: Practical machine learning tools and techniques.* 1st ed. NY: Morgan Kaufmann.

Zhang, C. & Ma, Y. e., 2012. *Ensemble machine learning: methods and applications.* 1st ed. NY: Springer Science & Business Media.

Zhang, Y. et al., 2015. Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Systems Journal,* 11(1), pp. 88-95.

Chen, Min, et al. "Disease prediction by machine learning over big data from healthcare communities." *Ieee Access* 5 (2017): 8869-8879.

<https://en.wikipedia.org/wiki/Disease>

<http://iosrjen.org/Papers/vol8_issue10/Version-3/I0810035359.pdf>

<https://www.python.org/doc/essays/blurb/>