IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.96145

Deep Learning Algorithm for Classification And Prediction of Lung Cancer

Pavitra B[1], Thanuja J C[2]

PG Student, Dept. of Master of Computer Application, Bangalore Institute of Technology, Bengaluru, Karnataka, India[1]

Assistant Professor, Dept. of Master of Computer Application, Bangalore Institute of Technology, Bengaluru, Karnataka, India[2]

Abstract: Diseases affect our daily lives as a result of most people's lifestyles. In the health-care field, most person's data is analysed as per the sickness that has impacted them as a result of their lifestyle, and some info is hidden that will be valuable in making good decisions. In this research, we use people's lifestyle disorders to process and analyse data using machine learning algorithms, and we use data visualization techniques to predict disease stage using machine learning algorithms Decision tree and K-Nearest Neighbour algorithm. And depending on that training data, we need to perform data validation to forecast the results. And with the dataset we can collect electronic medical records; all the medical record patient details are available as well as precision. The proposed model will help preserve the lives of the majority of patients, and we will be able to avoid the majority of incurable diseases if we can identify causes early on.

In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect Lung Cancer disease at early stage. This in turn will help to provide effective treatment to patients and avoid severe consequences.

Keywords: **Disease, machine learning, Lung Cancer, K-Nearest Neighbour algorithm**

I. INTRODUCTION

Despite the fact that medical data frequently contains hidden information, making decisions becomes more difficult. Machine learning is essential for detecting hidden patterns in medical data and analysing it. Machine learning is used to analyse data in various fields, including banking, transportation, government, marketing and healthcare. In the data mining discipline, machine learning deals with massive amounts of structured data. In medicine, machine learning is used for disease prediction, identification, and diagnosis. The primary goal of these methods is to diagnose Lung Cancer disease early, resulting in an earlier diagnosis and effective treatment of Lung Cancer disease.

Previously, medical studies on the extraction of information from a large number of structural data features were conducted. The three types of data found in a database are structured data, unstructured data, and semi-structured data. Structured data is data that includes real patient records such as testing records, EHRs, and data from medical test results, among many other things. The goal is to process a huge amount of data on Lung Cancer disease and, as an outcome, estimate the risk of Lung Cancer disease. Data cleaning as well as imputation are necessary when medical data is not in the correct format. Illness prediction is impossible due to poorly structured data, which can sometimes result in incorrect disease prediction. They had already finished with symptom-based disease prognosis. We used the nave bayes algorithm to forecast illness using structured data. In this paper, we perform the operation on medical structured data. The convolutional neural network is a deep learning concept that automatically extracts features from large datasets and produces the desired result. CNN was used to extract essential feature values from structured data in order to make illness predictions based on such dataset. The primary goal of this study is to predict Lung Cancer disease and also Lung Cancer disease risk using structured data. The second goal is to find the correct answer while dealing with missing values.

One of the hottest topics in the world of health statistics is determining the likelihood that a patient will become ill. When it comes to making use of a wide variety of health data, personalised prediction, which focuses on developing one-of-a-kind models for each individual patient rather than global models trained on the entire population, has proven to have advantages. In order to accurately capture individual characteristics, personalised prediction models use data from patient groups with comparable characteristics. An important step in the customised modelling process is accurately identifying and evaluating individual similarity based on previous records. Because a suitable vector

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.96145

representation is currently unavailable, electronic health records (EHRs) with an inconsistent sampling pattern and varying patient visit durations cannot be directly used to quantify patient similarity. In this paper, we present a novel time-fusion CNN framework for learning patient representations while also evaluating pairwise similarity. Deep learning techniques were used to create the framework. In contrast to a conventional CNN, our time fusion CNN can learn both local temporal correlations and contributions from each time period. The output data as from similarities learning approach is used to rate comparable patients, and this is identified as the probability distribution. We use similarity scores to predict individual diseases, and we investigate the implications of multiple available vector representations and similar learning metrics.

II. LITERATURE SURVEY

Health care generates massive amounts of data, which must be processed using specific techniques. Data processing is one of the techniques commonly used in machine learning algorithms, and cardiopathy is one of the leading causes of death worldwide. This machine learning algorithm technique predicts the emergence of cardiopathy. The results of these methods provide a percentage chance of developing cardiopathy. The datasets used are classified in terms of medical field attributes such as input and output data to split using the model selection method and this technique evaluates those parameters using the processing classification technique. Python programming is used to predict the output of the datasets.

Diabetes Mellitus or Diabetes has been compared to cancer and HIV (Human Immunodeficiency Virus). It rises when people have high glucose levels for an extended period of time. It has recently been identified as a risk factor for Alzheimer's disease, as well as a type one cause of blindness and nephritis. Disease prevention is also an important topic for research within the healthcare community.

According to World Health Organization research, heart disease is the leading cause of death worldwide. If this trend continues, by 2030, there will be approximately 23.4 million people who have died from cardiopathy. Due to the need for patient anonymity, the healthcare industry collects a large amount of data on cardiopathy, but in some cases, we will be unable to resolve the precise problem. An investigation into Principal Component Analysis was conducted during the research. Using supervised machine learning algorithms, this method finds the least amount number of attributes from which we can make a prediction. The goal of this study is to create supervised machine learning algorithms that can predict cardiopathy. To accomplish this goal, data processing has already been completed, and the raw data has been separated from the trained data. Finally, the final outcome will be predicted. Diabetes is a life-threatening disease in this day and age because it affects even small children. We can identify the specific symptoms and conduct predictions if we know how to use machine learning algorithms.

People today face a variety of diseases that are related to their daily activities, and it is difficult to predict whether the person will be affected in the early or late stages. As a result, accurate prediction is not possible for specific disease symptoms; it is a difficult task without the use of machine learning algorithms; and it is an upcoming challenge for people affected by specific diseases. Nowadays, science and technology have produced an infinite number of innovative ideas, and in some cases, we can use machine learning algorithms to predict various stages or the final level at which we can save a person's life.

III. EXISTING SYSTEM

Because of the environment and the choices people make regarding their way of life, people nowadays are prone to a wide variety of illnesses. Because of this, making an early diagnosis of a disease becomes critical importance. However, based just on symptoms, it could be challenging for medical practitioners to predict outcomes accurately.

Predicting when someone will get sick with accuracy is the biggest challenge. To help overcome this issue, disease prediction is a procedure that heavily relies on data mining. Because of the enormous volumes of data that are produced within the healthcare domains, early patient therapy has benefited from an accurate assessment of medical data. Data mining is the technique of leveraging knowledge of diseases to find patterns in vast amounts of medical data that have never been seen before. We suggested a strategy for drawing conclusions about the patient's general health based on the symptoms.

Making an accurate diagnosis of a patient through the use of clinical analysis and evaluation is crucial. It's feasible that using computer-driven decision support systems will become absolutely vital for making critical decisions. The health

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.96145

care industry generates a sizable volume of information about clinical assessments, patient reports, treatments, medicine, follow-up visits and other subjects. It takes a tremendous amount of planning to do it right.

The quality of the data association has decreased to an undesirable degree as a result of poor information management. A valid method that can concentrate and processing data in a way that is both practical and efficient is urgently needed due to the increase in data volume. The task of creating accurate sickness forecasts is growing in importance due to exponential growth rate of medical data. On the other hand, processing vast amounts of data is a very taxing task, making it very challenging to predict a patient's disease.

IV. PROPOSED SYSTEM

Within the context of this study, we propose that the use of the K-Nearest Neighbor (KNN) and Decision Tree (CNN) machine learning algorithms results in good disease prediction. Both the prediction and diagnosis of a condition require a set of symptoms. This general disease prediction considers the person's lifestyle decisions as well as the data from their check-ups in order to produce an accurate forecast. The decision tree approach has a disease prediction accuracy of 84.5 percent, which is higher than the KNN method's illness prediction accuracy. Depending on the severity of the danger, this system can anticipate general disease and then assess whether the risk of general disease is high or low.

The proposed method builds classifiers that can categorise data based on the characteristics of that data by utilising a wide range of machine learning techniques. The data set has been divided into two or more separate groups. Such classifiers are used to analyse medical data and predict illnesses. Machine learning is currently used in so many different fields that it is totally possible to utilise it every day without even realising it. In contrast to other machine learning algorithms that can only effectively use structured data and take a long time to compute, CNN uses both structured and unstructured data to identify hospitals. Additionally, they save all the data in the form of a training dataset, and they use complicated calculation algorithms, which makes them lazy.

V. RESULT

The following figure shows the result after analyzing the patient data from all the reports and gives the final report of analysis which predicts the level of cancer risk from the range of symptoms of patient.

Fig 5.1: Report Page

VICONCLUSION

Diseases that are related to how a person or group of people live their lives, such as their habits, activities, and diet, for example. In the EMR data, we can also obtain private information about wealthy individuals. This proposed system would use ML algorithm techniques such as Decision Trees and K-Nearest Neighbor to predict diseases in their early or late stages and recover the patient as quickly as possible. In addition, we developed a machine learning model based on EMR data that examines patient input data and compares it to learned datasets that represent particular diseases, enabling us to stop disease progression. People's bad eating habits, stress at work, and lack of physical activity are the

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.96145

main causes of most diseases. The majority of patients' lives will be preserved by the suggested methods, and if the causes can be found early enough, most terminal diseases can be avoided.

REFERENCES

[1] Anand, A. and Shakti, D., 2015. Prediction of diabetes based on personal lifestyle indicators. In Next generation computing technologies (NGCT), 2015 1st international conference on (pp. 673–676). IEEE.
[2] Kanchan, B.D. and Kishor, M.M., 2016. Study of machine learning algorithms for special disease prediction using principal of component analysis. In Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016 International Conference on (pp. 5–10). IEEE.
[3]. Sayali Ambekar and Dr.Rashmi Phalnikar, 2018. Disease prediction by using machine learning, International Journal of Computer Engineering and Applications, vol. 12, pp. 1–6.
[4]. Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H. and Jahan, H., 2018. PRMT: Predicting Risk Factor of Obesity among MiddleAged People Using Data Mining Techniques. Procedia Computer Science, 132, pp. 1068–1076.
[5]. Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H. and Jahan, H., 2018. PRMT: Predicting Risk Factor of Obesity among MiddleAged People Using Data Mining Techniques. Procedia Computer Science, 132, pp. 1068–1076.
[6]. Mishra, A.K., Keserwani, P.K., Samaddar, S.G., Lamichaney, H.B. and Mishra, A.K., 2018. A decision support system in healthcare prediction. In Advanced Computational and Communication Paradigms (pp. 156–167). Springer, Singapore.