Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/46

Developing a Machine Learning-Based Framework for Disease Prediction

Kumud Pant[1], Dibyahash Bordoloi[2], Bhasker Pant[3]

[1]Department of Biotechnology,Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002
[2]Head of the Department, Department of Computer Science & Engineering,Graphic Era Hill University, Dehradun, Uttarakhand India, 248002
[3]Department of Computer Science & Engineering,Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002

ABSTRACT

Due to environmental conditions and contemporary lifestyles, individuals encounter a variety of ailments presently. Predicting diseases at an early stage thus becomes an essential endeavor. However, the precise prognosis from symptoms becomes increasingly challenging for doctors. Correctly predicting sickness is the most difficult job. To solve this issue, data mining plays a crucial role in predicting the illness. The health sector generates an increasing quantity of data each year. Due to the rise in the quantity of information available in the healthcare field, initial patient treatment has benefited from the thorough assessment of medical data. Massive amounts of patient records are mined for hidden patterns with the use of disease-specific data. We presented a broad illness prediction system depending on the patient's health condition. We employ the K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) machine learning algorithms for effective illness prediction. Knowledge of illness symptoms is essential for prediction and diagnosis. In this generalized illness prediction, the individual's lifestyle and test results are considered for an effective prediction outcome. CNN's general illness accuracy rate is 89.3%, that is higher than the KNN approach. Additionally, KNN is more demanding in terms of time and storage space than CNN. Following general illness forecasting, this method can provide the risk linked to the overall disease, whether that risk is low or high.

Keywords: Machine Learning approaches, Disease Forecasting, Healthcare sector, KNN, CNN.

INTRODUCTION

The development of Artificial Intelligence (AI) has made computers more clever and given them the ability to think for themselves. Researchers in artificial intelligence often include machine learning (ML) as a subfield in their investigations. There is a consensus among analysts that one cannot get insight without first gaining knowledge[1]. Many machine learning algorithms are available, including unsupervised, semi-supervised, supervised learning, deep learning, evolutionary, and reinforcement learning. These discoveries are put using to categorize massive amounts of data

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/46

rapidly. Therefore, we apply the K-nearest Neighbor (KNN) and Convolutional neural network (CNN) algorithms for precise illness predictions and rapid categorization of large amounts of data. Data mining performs a vital task, and the categorization of massive data employing ML becomes so much easier because healthcare information is rising daily, making it an essential task to utilize that data to forecast diseases accurately. However, handling large amounts of data is also extremely crucial in a broad sense.

It is essential to have a solid understanding of the definitive diagnosis that should be given to individuals based on their clinical assessment and examination. For persuasive determination, computer-based decision support systems may take on a vital role. The health care sector generates a large amount of data on clinical evaluations, reports about patients, treatments, follow-up appointments, medicines, and other topics [2]. It is difficult to choreograph everything optimally. Because of carelessness in handling the content, the data quality has suffered as a result. Any increase in the amount of data necessitates the development of a reliable method to focus, manage, and analyze the data in an effective and time-saving manner. To design a classifier that is capable of separating the data based on their properties, one of the various programs available for ML is used. The dataset is divided up into a total of two or more different categories [3]. These classifiers are applied for the examination of medical information as well as the forecasting of sickness.

ML is so pervasive in today's world that even those unaware of its existence are likely to use it daily. CNN's categorization process utilizes a clinic's structured and unstructured data [4]. Whereas other ML approaches can only be used with structured data, the computations take a long time; they are inefficient since they keep all the records as training data, and they employ a complicated approach to compute their results. Numerous publically accessible datasets are available for research purposes, such as disease datasets utilized in this work [5], [11]. Section I explains the general illness prediction utilizing classification methods like KNN and CNN [5]. The literature assessment of previously implemented systems is presented in Section II, and the proposed system deployment specifics are outlined in Section III. The experimental research, the findings, and a discussion of the proposed methodology are presented in Section IV. The conclusion of our proposed approach may be found in Section V. While the list of references for the work is supplied at the very end.

Related Works

A novel CNN-based multimodal illness predictive system is described in Reference [6]. This algorithm makes use of both structured and unstructured data from hospitals. The authors developed a disease prediction system that may be used for various geographic areas. They conducted illness prediction research on three different diseases, including diabetes, heart disease, and cerebral infarction. The analysis of the structured data allows for accurate illness prognosis. Various ML algorithms, such as naive bayes, decision tree, and KNN algorithm, are used to make accurate predictions of cardiovascular disease, hypertension, and cerebral infarction. The decision tree algorithm results are superior to those of the KNN method and the Naive Bayes algorithm. In addition, they can forecast whether or not a patient would suffer from an elevated danger of cerebral ischemia or a reduced risk of hemorrhagic stroke. They applied CNNs to perform multimodel illness risk prediction on text data to provide an accurate prediction of the risk of cerebral infarction. The CNN-based unimodal illness risk prediction algorithm is put up against the CNN-based multimodel

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/46

disease risk prediction algorithm in order to determine whether one has a higher level of accuracy. With this method, the accuracy of illness prediction can reach up to 95.2%, and it can do so quicker than previous CNN-based unimodal disease risk prediction algorithms. The stages of the CNN-UDRP method and those of the CNN-based multimodel illness prediction method are identical; the only difference is that the assessment stages of the CNN-based multimodel illness classification techniques include two extra steps. This research works on unstructured and structured datasets, which are examples of different types of datasets. The author dealt with data that was not structured. Previous work solely relied on structured information; however, neither of the authors has done any work using unstructured or semi-structured data. However, the success of this work relies on both organized and unstructured data.

The authors of reference [7] created the Alzheimer's disease risk prediction system using the EHR information provided by the patient as their primary source of information. In this instance, they used an active learning setting to find a solution to a genuine issue that the patient was experiencing. This led to the construction of the effective patient risk model. A predictive risk model is applied to determine the likelihood of Alzheimer's disease. The authors of Reference [8] envisaged wearables 2.0 platform wherein smart washable clothes would be designed to enhance the quality of treatment and the patient experience inside the next-generation medical system. The authors conceived of a new data collecting method based on the Internet of Things, in that newly designed sensor-based, reusable, and intelligent fabric. The doctor was able to record the participant's physiological parameters with the usage of this fabric. And considering that, with the assistance of the physiological parameters, additional data were analyzed. With the assistance of this element, the customer is ready to accumulate the patient's physiological situation in addition to providing information regarding the participant's mental wellbeing condition through a cloud-based framework. This reusable smart fabric asymmetry primarily consists of sensors, cables, and electrodes. This fabric was used to record the patient's vitals in order to assess their health. In addition, this data is utilized for the purpose of the data study. Presented and discussed the challenges that arose in the process of designing the portable 2.0 framework. The problems with the current system include collecting physiological parameters, experiencing terrible psychological impacts, being anti-wireless for skin area networking, and collecting a large number of physiological parameters sustainably, among other problems. Many actions are carried out on files, such as data analysis, surveillance, and forecasting. The author categorizes the smart textiles' core elements that represent Wearables 2.0 into the following groups: sensor integration, electrical-cable-based connectivity, and digital components. Many different possibilities are mentioned in this article, such as tracking chronic diseases, caring for old persons, caring for emotions, etc.

Later in [9], the authors created cloud-based healthcare –Cps platform that controls the enormous volume of data that pertains to biomedicine. They spoke about expanding the vast amounts of data in the medical area. The data is made in a shorter length of time than usual, and the characteristics of the data are saved in various formats; hence, the issue was with the massive amounts of data. One of these technologies, cloud computing, and the other, big data technology, were prioritized in developing the healthcare information system built here. This system carried out a variety of tasks, including analysis, monitoring, and prediction of data, all in a cloud-like environment. A person may get further knowledge on how to handle and manage the situation with the assistance of this

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/46

system. Cloud storage hosts a massive volume of biological data. The data collecting layer, the data management layer, and the data-oriented layer make up the three levels that are considered in the system. The data collecting layer was responsible for storing the information in the specific standard format. The layer of data management is utilized for parallel computing and distributed storage. With this system's assistance, several different procedures may be carried out. Health-cps system. In addition, the many different healthcare-related services are known by this system.

In [10], the authors examined how to manage a significant volume of medical data via the cloud and presented a telemedicine platform. The author of this research suggested improvements to the telehealth network, which is mostly focused on data exchange among all cloud-based telehealth. However, there are other problems with cloud data sharing, including network bandwidth and virtual machine switching. A cloud-based solution to data sharing is recommended for improved data sharing using data sharing ideas. Here is a design for the best way to a paradigm for sharing telehealth. With the use of this model, the author focuses on temporal restrictions, network capabilities, and transmission probability. For this, the author created a brand-new, ideal method for exchanging massive data. Users are provided with the best method for processing biological data by this algorithm. The authors suggested a top clinical decision-making system that uses patient past data to forecast the illness. Numerous illnesses and an unanticipated pattern of patient state are predicted in this and created a top-notch diagnostic decision-making tool that is utilized for precise illness prognosis based on historical information. Several illnesses were also identified in it, employing ideas and unknown patterns. Additionally, 2D/3D graphs and pie charts are used to visualize data.

A comparison of several ML approaches, including fuzzy logic, fuzzy neural networks, and decision trees, is provided in reference [12]. They use the liver dataset to categorize and conduct comprehensive studies. Based on research, Fuzzy Neutral Network performs 92% more accurately than ML algorithms in classifying liver illness datasets. The author is skilled at doing classification extremely well and providing very good performances. The author employed Simplified Fuzzy ARTMAP in a variety of application areas. The author has concluded that on the supplied data set, machine learning techniques like Naive Bayes and Apriori [13] are very helpful for illness detection. Here, small-volume data, such as symptoms or prior information gleaned from the medical diagnostic, are employed for forecasting. Limitations to this study also couldn't take into account a vast dataset. Classifying the expanding amount of healthcare data that is needed nowadays is difficult.

In [14], the authors suggested using a CNN-MDRP method to predict diseases using a large amount of structured and unstructured clinical information. Utilizing Neavi-Bayes or the CNNUDRP existent method employs just structured data. Still, CNN-MDRP concentrates on structured and unstructured data, resulting in higher illness prediction accuracy and faster forecasting times than CNNUDRP.

Proposed System

At first, we get a disease dataset in the format of a symptoms checklist from the UCI machine-learning portal. Afterward, the dataset undergoes preprocessing, eliminating unwanted characters

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/46

like commas, punctuation, and whitespace. And then, we utilize that information as a training set. Next, key features are picked and retrieved features. The information is then put into categories employing KNN and CNN classification methods. Reliable illness forecasts are now possible using ML. Figure 1 depicts the overall architecture of this research work.

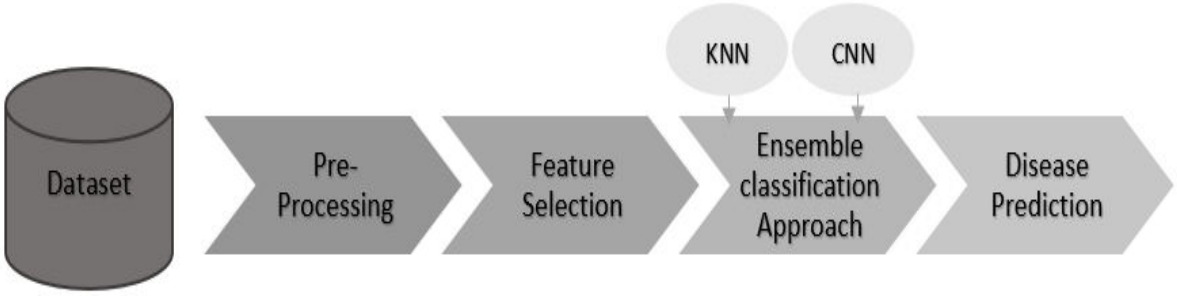

Figure 1: Proposed Architecture

The first process involves transforming the dataset into a vector format. Second, we use a process called word embedding, that uses dummy data points of zero whenever possible. The convolutional layer is the final product of word embedding. Third, we use the output from the convolution layer to feed to the pooling layer, where we do the max pooling function. The fourth stage is transforming the dataset into a fixed-length vector format using Max pooling. A fully connected neural network is linked to the pooling layer. Fifth, a complete connection network is linked to the softmax classifier.

Result Evaluation

The entire experiment was programmed in Java in conjunction with Netbeans tools and MySql as the database. Techniques and approaches, as well as alternative classification techniques and other feature extraction techniques, are also programmed in Java and executed on a computer with the setup of an Intel Core i5-6200U, 2.30 GHz Windows 10 (64-bit) workstation. This study made use of a patient illness dataset [5, 11] obtained from the UCI ML portal.

This section compares and contrasts the efficiency and precision of the KNN and CNN classification methods. There is a comparison between the KNN and CNN methods' accuracy at different thresholds shown in Table 1. Figure2 depicts this relationship graphically, with the algorithm shown along the X axis and the percentage of accuracy displayed along the Y axis. Disease prediction using CNN is more reliable than using KNN.

Table 1: Comparison of Accuracy %

Algorithm	*Accuracy (%)*
KNN	92.3
CNN	97.5

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/46

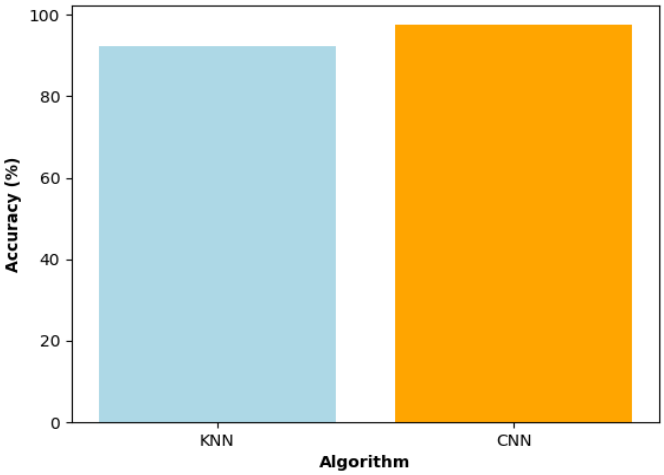

Figure 2: Accuracy Comparison

As can be seen in Table 2, both the KNN and CNN algorithms take a considerable amount of time, regardless of their respective sizes. See Figure 3 for a graphic representation of this; the X-axis represents algorithms, and the Y-axis represents time in milliseconds. For more extensive dataset classification, CNN is faster than KNN.

Table 2: Computational Time

Algorithm	*Computational time (ms)*
KNN	13521
CNN	11233

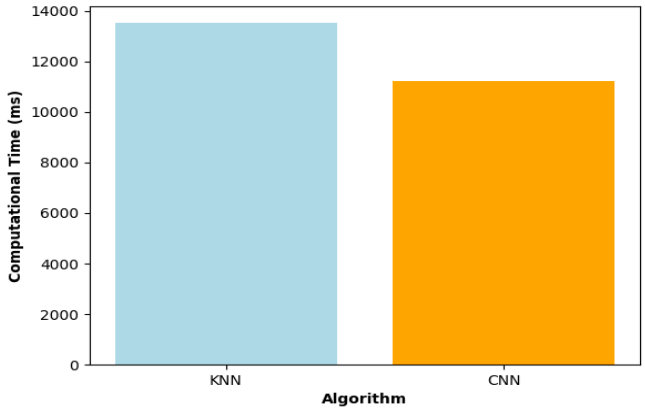

Figure 3: Computational Time comparison

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/46

Conclusion

This paper proposed a broad supervised ML algorithm-based illness predictive model. We used KNN and CNN techniques to categorize patient information since medical information is expanding rapidly and has to be processed so that specific diseases may be predicted from symptoms. By providing the input of medical files, which aid in understanding the degree of illness prediction, we could produce effective generalized illness prediction models. Due to this technique, illness and risk prognosis may be accomplished with little effort and expense. We evaluate the outcomes of the KNN and CNN algorithms in regard to precision and processing time. The accuracy of the CNN model is higher than that of the KNN algorithm, and CNN's processing time is lower than that of KNN. In terms of precision and timing, we may conclude that CNN is superior to KNN.

References

1. B. Nithya , Dr. V. Ilango Professor, "Predictive Analytics in Health Care Using Machine Learning Tools and Techniques," *International Conference on Intelligent Computing and Control Systems*,2017
2. Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "Data Mining and Visualization for prediction of multiple Diseases in Healthcare," *in IEEE big data analytics and computational intelligence*, Oct 2017 pp.2325.
3. Shanthi Mendis, Pekka Puska, Bo Norrving, World Health Organization (2011), Global Atlas on Cardiovascular Disease Prevention and Control, PP. 3– 18. ISBN 978-92-4-156437-3.
4. Amin, S.U.; Agarwal, K.; Beg, R., "Genetic neural network based data mining in prediction of heart disease using risk factors*", IEEE Conference on Information & Communication Technologies (ICT),* vol., no.,pp.1227-31,11-12 April 2013
5. Heart disease Dataset-*WWW.UCI Repository. Com*
6. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", ," *IEEE Access,* vol. 5, no. 1, pp. 8869–8879, 2017.
7. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Springer Data Mining Knowl. Discovery,* vol. 29, no. 4, pp. 1070–1093, 2015.
8. IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system," *IEEE Commun*. , vol. 55, no. 1, pp. 54–61, Jan. 2017.
9. Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," *IEEE Syst.* J., vol. 11, no. 1, pp. 88–95, Mar. 2017.
10. L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing," *in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud),* Nov. 2016, pp. 184–189.
11. Disease and symptoms dataset –*www.github.com.*
12. S.Leoni Sharmila, C.Dharuman and P.Venkatesan "Disease Classification Using Machine Learning Algorithms -A Comparative Study", *International Journal of Pure and Applied Mathematics* Volume 114 No. 6 2017, 1-10

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/46

13. Allen Daniel Sunny1, Sajal Kulshreshtha, Satyam Singh3, Srinabh, Mr. Mohan Ba, Dr. Sarojadevi H "Disease Diagnosis System By Exploring Machine Learning Algorithms", *International Journal of Innovations in Engineering and Technology (IJIET)* Volume 10 Issue 2 May 2018.
14. Shraddha Subhash Shirsath "Disease Prediction Using Machine Learning Over Big Data" International Journal of Innovative Research in Science, Vol. 7, Issue 6, June 2018