BIGDATA ANALYTICS FOR HEALTHCARE USING ARTIFICIAL NEURAL NETWORK

Dr.G.Saravanan AP/CSE, Divyaabharathi.R, Gobi suresh.B, Jhagavin.T,
Dept of CSE, Erode Sengunthar Engineering College,

Erode, Tamil Nadu

ABSTRACT: Chronic airway inflammation caused by diseases is known to cause episodic wheezing, throat tightness, coughing, and shortness of breath. In this study, a machine learning-based algorithm for predicting illness risk is presented (ML). From the provided data set, this project determines a link between the symptoms and prognosis. The link between the indoor PM and weather data is mapped to the discovered values using a (CNN) architecture. The root mean square and mean absolute error accuracy measures of the suggested method are compared to those of cutting-edge deep neural network (DNN)-based techniques. Additionally, the accuracy of the classification methods K-Nearest Neighbor and Support Vector Machine are carried out. The SVM, KNN, and CNN classification techniques used in the new data set help to better accurately predict the illnesses category. Python 3.7 is the coding language employed.

Keywords: Machine Learning**,** Support Vector Machine, K-Nearest Neighbor, Convolutional Neural Network.

I. INTRODUCTION

Big data is important in every industry, including the health care sector, throughout the world. It changes how carefully to manage doctors and their patients. We could anticipate more precise outcomes and insights for the health care businesses from the large sample size of data. Like many other industries, the health care sector is made up of heterogeneous, interconnected sectors that are difficult to manage accurately while yet meeting patient demands for higher quality care at lower costs.

Emerging technologies are being gradually incorporated into the healthcare sector, where big data analytics play a crucial role in providing hospitals and patients with useful business insights. In the technical world, data analysis is crucial in any industry when the amount of data is so small. But the big data era has arrived in the modern world. According to current statistics, data analytics will become increasingly significant in the operational, clinical, and banking/financial sectors of the health care industry.

Government and public organizations may use the gathered data to develop or improve procedures, regulations, and trainings. Overall, the project has the ability to increase awareness of the need to provide the finest care possible in every healthcare setting.

The majority of patients lack education and are unfamiliar with exact treatments. The majority of patients/people go to private health care facilities, yet these facilities are unable to save information about patients and their illnesses. Therefore, it is necessary to plan health fairs that inform and raise awareness among the locals. This study provides information on diagnosis and various health risks.

The remainder of this essay is structured as follows: Section 2 highlights earlier efforts and their shortcomings while reviewing the current methodologies under recent investigations. The study's proposed approach is presented in Section 3. Findings are presented in Section 4, and the study is concluded in Section 5.

II. LITERATURE REVIEW

The term "big data" refers to the increasing amounts of organized, unstructured, and semi-structured data that have been produced by several organizations worldwide in recent years. The demand to maintain the big data being produced by many sources, which are recognized for producing enormous volumes of heterogeneous data, has confronted the health industry sectors.

To manage the enormous amounts of data in the healthcare sectors, more and more big-data analytics technologies and methodologies are being developed. The authors of this study explored the effects of big data on healthcare and the many tools available in the Hadoop ecosystem to handle it.

They also looked at the conceptual framework for big data analytics in the healthcare industry, which includes the genomic database, text/imagery, clinical DSS, and the history of data collection from various branches (decisions support system).

Every day, data is produced by a variety of applications and geographical research projects for a variety of reasons, including crime detection, catastrophe evaluation, weather forecasting, and the health business, to mention a few. Today's big data scenarios are linked to fundamental technologies and a wide range of businesses, like Facebook, Google, and IBM, which mine enormous volumes of amassed data for useful information [3-5].

Healthcare is about to enter an era of open data. In all industries, including healthcare, big data are being produced quickly in relation to patient compliance, care, and other regulatory needs. Treatment delivery strategies are changing quickly as the world's population and average lifespan both increase [6]. Some of the decisions underlying these quick changes are based on data. Healthcare investors promised new information from big data, so named for both its amount and variety and complexity.

Experts and investors in the pharmaceutical sector have started systematically reviewing and analyzing big data to get insights, but these efforts are still in their infancy and need to be integrated in order to address issues with healthcare delivery and improve the standard of care. Early big data analytics systems for health care informatics are set up in a variety of scenarios, such as the analysis of patient characteristics/determination of treatment cost and outcomes to identify the best and most economical therapies [6]. Health informatics is the study of healthcare information through the integration of health care sciences, computing sciences, and information sciences. Data storage, acquisition, and retrieval are all part of health informatics, which helps medical professionals deliver better treatment.

They came to the conclusion that they had given a thorough explanation of big data in general as well as how it affects healthcare systems, and that big data has a major impact on both the health care system and on healthcare in general. In addition, they suggested using conceptual architecture and Hadoop terminologies to solve health care problems involving big data. This would entail using big data produced by various levels of medical data as well as developing analytical techniques for this data in order to find answers to medical questionsBig data and health care analytics work together to create treatments for specific patients that are more effective than those that are helpful for the majority of people. This allows doctors to prescribe the right meds for each patient individually.

As we all know, big data analytics is still in its infancy, and current tools and techniques are unable to address big data-related issues. Massive data is thought of as big systems that provide enormous hazards and challenges. Therefore, a lot of study is needed to find solutions to the issues the healthcare system is currently facing.

The authors of the publication [2] provided a succinct overview of big data and its use in health care applications. The management of data expansion in the health care industries is evidently being helped by the usage of big data architecture and associated methodologies. Here, an empirical study is first done to examine the functions of big data in the health care sector.

It has been noted that much work has been done on big data in the healthcare industries. Currently, it is difficult to imagine how big data and machine learning will affect the health care industries. It has been noted that the majority of authors that utilize big data analytics and machine learning to diagnose diseases do not place a high priority on data security and privacy.

For handling big data for the medical sectors, a novel design of smart/secure health information system using machine learning and also the enhanced security mechanism is provided. Incorporating the best storage options along with a data security layer that protects both security and privacy constitutes innovation. The incorporation of various methods includes end-point validation, dynamic data encryption, masking encryption, granular access control, and activity monitoring. The hybrid 4-layer health care approach that has been presented appears to be more effective at diagnosing diseases using big data.

Today, all firms are experiencing uncontrollable data growth due to the quick expansion of internet cloud computing [7]. Wal-Mart imports 2.5 petabytes of data into databases every hour or so, Facebook processes more than 250 million photographs and 900 million objects daily, etc. [8].

Explanations are necessary in order to extract useful insights from the datasets as a result of the data's fast expansion. Effective data use is crucial since it is thought of as the foundation of an organization. The diagnosis of diseases, forecasting of sales, economic analysis, social network analysis, and corporate management, among other things, all benefit greatly from effective data analysis.

Some businesses employ reports that contain organized data from analytics. The concept of big data was first presented in the 1999 publication "Visually Exploring Gigabyte Datasets in Real Time"

[9]. In his 2001 work 3D Data Management: Controlling Data Volume, Velocity, and Variety, Doug Laney defined the 3Vs, or velocity, volume, and variety, as the characteristics of big data.

Hadoop is one of the most well-liked tools for managing and processing large amounts of unstructured data. Big data is a term used to describe massive, complex amounts of data that are difficult to handle using traditional methods and are obtained from a number of sources, including the web, mobile devices, office applications, and digital repositories.

Big data is the act of storing and managing data for future analysis. It is not just about enormous data sizes. Computing therefore enmeshes data with more volume, diversity, and velocity as people become more digital. Big data in the healthcare sector presents enormous prospects for enhancing efficiency and quality in healthcare, identifying health hazards, controlling human health by early disease diagnosis, and assisting in better decision-making.

They came to the conclusion that they had given a thorough explanation of big data in general and in the health care system, as well as a concise overview. Big data, they said, plays a significant role in health care informatics and has a significant impact on the health care system and big data's four Vs.

They also suggested using the conceptual architecture to address health care issues in the context of big data using Hadoop terminologies. This involved using big data, which was produced by different levels of medical data, as well as developments for both analyzing this data and obtaining answers to medical questions.

III. PROPOSED METHODOLOGY

To identify disease in the provided data set, the current system uses the Nave Bayes classification technique. Kaggle is where the dataset was obtained. In the current system, preprocessing like null value eradication is not processed. When classifying, every feature is collected, which lengthens the process. Confusion matrix is created using the calculation of accuracy scores. The disadvantages are:

- NBS Classification is not given more accuracy for given new test data.
- Feature reduction before classification is not carried out.
- Decision Tree takes more time if the data set size is growing
- Data columns with numeric values only take for NBS classification.

Along with current system algorithms, the proposed system focuses on SVM classification techniques. The dataset was downloaded from Kaggle and preprocessed, including the deletion of null values. For better classification, key traits are chosen. Calculated and printed is the accuracy score. Accuracy score is used to calculate the confusion matrix as well. The suggested application includes the components listed below.

1. **DATA SET COLLECTION**
2. **DATA SET SUBSETTING**

3. **NBC CLASSIFICATION**

4. **DT CLASSIFICATION**

5. **SVM CLASSIFICATION**

1. DATA SET COLLECTION

The dataset is saved as records in a single Excel worksheet and includes columns for "itching skin rash," "nodal skin eruptions," "constant sneezing," "shivering and chills," and other symptoms with "prognosis" as the disease name. This serves as the project's input.

2. DATA SET SUBSETTING

The dataset is saved as records in a single Excel worksheet and includes columns for "itching skin rash," "nodal skin eruptions," "constant sneezing," "shivering and chills," and other symptoms with "prognosis" as the disease name. This serves as the project's input, of which 4920 (total; 120 records for each disease) are used for training records and 15% for testing records, which are divided and offered for classifications.

3. NBC CLASSIFICATION

This module employs Naive Bayes Classification, wherein 15% of the data are used as test data and 85% of the data in the given data set are used as training data. Numerical values are generated from the text (categorical) columns. The model is then forecasted using test data after being trained with training data. Of which, the majority of illnesses are categorized as having itching present or not.

4. DT CLASSIFICATION

In this module, a decision tree classification method is utilized, in which 15% of the data are used as test data and 85% of the data in the given data set are used as training data. Numerical values are generated from the text (categorical) columns. The model is then forecasted using test data after being trained with training data. Of which, the majority of illnesses are categorized as having itching present or not.

5. SVM CLASSIFICATION

85% of the data in the given data set are used as training data and 15% are utilized as test data in this module's Support Vector Machine Based Classification. Numerical values are generated from the text (categorical) columns.

IV. FINDINGS

- SVM classification is considered suitable for the given new test data.
- SVM classification gives more accuracy.
- SVM supports well even if the dataset size is big.
- SVM based prediction model is worked out to find algorithm efficiency with different test data sizes.

V. CONCLUSION

Since it takes more experience and needs the use of medical testing, identifying potential diseases in patients is a difficult task for clinical personnel and researchers. The study identifies the best classification algorithm that can enhance accuracy when classifying normal and abnormal people. It includes the calculation of the accuracy scores for Support Vector Machine, Naive Bayes, and decision tree classification. In the new data set, the application of SVM, NBS, and DT classification aids in more accurate disease prediction.

REFERENCES

[1] Prableen Kaur, Manik Sharma, Mamta Mittal,Big Data and Machine Learning Based Secure Healthcare Framework, Procedia Computer Science 132

(2018) 1049–1059.

[2] Sunil Kumar, Maninder Singh, Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools, BIG DATA MINING AND ANALYTICS ISSN222096-0654, 05/06, pp48–57, Volume 2, Number 1, March 2019, DOI: 10.26599/BDMA.2018.9020031

[3] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods and analytics, International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015.

[4] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "Big Data", Hadoop and cloud computing in genomics, Journal of Biomedical Informatics, vol. 46, no. 5, pp. 774–781, 2013.

[5] C. L. P. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, vol. 275, pp. 314–347, 2014.

[6] M. Herland, T. M. Khoshgoftaar, and R.Wald, A review of data mining using big data in health informatics, Journal of Big Data, vol. 1, no. 1, p. 2, 2014.

[7] Ozgur C., Kleckner M. and Li Y. (2015) "Selection of

Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities." Sage Open: 1-12.

[8] Sagiroglu S. and Sinanc D. (2013) "Big Data: A

Review. Presented in International Conference: Collaboration Technologies and Systems (CTS)." IEEE Xplore.

[9] Picciano A. G. (2012) "The Evolution of Big Data and Learning Analytics in American Higher Education."

Journal of Asynchronous Learning Networks 16(3): 9-20.