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 forboth 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 helpsmedical 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 is observed that significant works have been done for big data in health care sectors. Nowadays, it is intricate for envisioning the way machine learning as well as big data influences the health care industries. It has been observed that most authors implemented the machine learning and big data analytics use in disease diagnosis are not giving significant weightage to data privacy and security.

Here, a new design of smart/secure health care information system with the use of machine learning

and also the advanced security mechanism are proposed for handling big data for medical industries. The innovation lies in incorporation of optimal storage as well as data security layer used for maintaining both security and privacy. Various techniques suchas activity monitoring, masking encryption, granular access control, dynamic data encryption and end-point validation are incorporated. The proposed hybrid 4 layer health care model seems to be more effective in disease diagnostic big data system.

Today, due to expeditious development of internet cloud computing, data grows fastly at uncontrollable rate in all organizations [7]. Wal-Marts imports 2.5 petabytes of data approximately every hour in to databases, Facebook handles more than two fifty million photos and nine hundred million objects each and every day etc [8].

Due to this explosive growth of data, explications are must to glean valuable insights from the datasets. The effective data utilization is important as it is deemed as the building blocks for an organization. The effective data analysis are very useful in the disease diagnosis, sale forecasting, economic analysis, social network analysis and business management, etc.

Some organizations use formerly analytics in the organized data in form of reports. The early idea of big data were introduced in the paper "Visually Exploring Gigabyte Datasets in Real Time" and were published in 1999 [9]. In 2001, Doug Laney defines the big data characteristics in 3V's i.e. Velocity, Volume and Variety in the paper 3D Data Management: Controlling Data Volume, Velocity and Variety.

Hadoop is one of the most dominant frameworks that is used for managing and analyzing the unstructured big data. In general, the big data refers to voluminous and complex amount of data collected from various sources suchas web, mobile devices, enterprise applications and digital repositories that can not be easily managed with the use of traditional tools.

Big data is not only about large data size, but it is an act of storing and managing data for eventual analysis. As the persons are getting digitized, therefore, computing embroils data with greater volume, variety, and velocity. Big data offer tremendous opportunities in health care sector for improving efficiency and quality in health care, health threats detection, managing human health by diagnosis disease in early stage and in assisting better decisions.

They concluded that they have provided an in- depth description and brief overview of th big data in general and in health care system, which plays a significant role in health care informatics and influences greatly the health care system and big data four Vs in health care.

They also proposed use of the conceptual architecture to solve health care problems in big data with Hadoop terminologies, that involved the utilization of big data, generated by various levels of medical data and the developments for analyzing this data and also for obtaining answers to medical questions.

III. PROPOSED METHODOLOGY

The existing system focuses on Naïve Bayes classification algorithm to detect disease among the given data set. The dataset is taken from kaggle. Preprocessing such as null value elimination is not processing in existing system. All features are taken during classification which increases time. Confusion matrix is prepared with accuracy score calculation. Thedrawbacks are:

- NBS Classification is not given moreaccuracy for given new test data.
- Feature reduction before classification isnot carried out.
- Decision Tree takes more time if the dataset size is growing
- Data columns with numeric values onlytake for NBS classification.

The proposed system focuses on SVM classification algorithms along with existing system algorithms. The dataset is taken from kaggle and preprocessed such as null value elimination. Important features are selected for better classification. Accuracy Score is calculated and printed. Confusion matrix is also calculated with accuracy score. The following modules are present in the proposed application.

1. **DATA SET COLLECTION**
2. **DATA SET SUBSETTING**
3. **NBC CLASSIFICATION**
4. **DT CLASSIFICATION**
5. **SVM CLASSIFICATION**

1. DATA SET COLLECTION

The dataset which contains columns (itching skin_rash, nodal_skin_eruptions, continuous sneezing, shivering, chills, etc with prognosis as disease name) are saved in a single Excel workbook as records. This is the input for the project.

2. DATA SET SUBSETTING

The dataset which contains columns (itching skin_rash, nodal_skin_eruptions, continuous sneezing, shivering, chills, etc with prognosis as disease name) are saved in a single Excel workbook as records. This is the input for the project in which 4920 (collectively (120 records for each disease) for training records and 15% of which is taken for testing records are split andgiven for classifications.

3. NBC CLASSIFICATION

In this module, Naive Bayes Classification is used in which 85% of the data in given data set is taken as training data and 15% of the data is taken as test data. The text (categorical) columns are converted into numerical values. Then the model is trained with training data and then predicted with test data. Of which, most of the disease are classified as itching present or not.

4. DT CLASSIFICATION

In this module, Decision Tree Classification is used in which 85% of the data in given data set is takenas training data and 15% of the data is taken as test data. The text (categorical) columns are converted into numerical values. Then the model is trained with training data and then predicted with test data. Of which, most of the disease are classified as itching present or not.

5. SVM CLASSIFICATION

In this module, Support Vector Machine Based Classification is used in which 85% of the data in given data set is taken as training data and 15% of the data is taken as test data. The text (categorical) columns are converted into numerical values. Then the model is trained with training data and then predicted with test data. Of which, most of the disease are classified as itching present or not.

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 sizeis big.
- SVM based prediction model is worked out to find algorithm efficiency with different test data sizes.

V. CONCLUSION

Since the finding of possibility of diseases in patients is a tough task for clinical persons and researchers as it requires more experience and medical tests need to be taken. The project finds the best classification algorithm which is suitable for providing accuracy improvement during classification of normal and abnormal persons. It contains Support vector machine, Naïve Bayes and decision tree classification with their accuracy score calculation. The applied SVM, NBS, DT classification helps for predicting the disease with higher accuracy in the new data set.

REFERENCES

[1] Prableen Kaur, Manik Sharma, Mamta Mittal,Big Dataand 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 ofBiomedical 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 Guidefor 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." Journalof Asynchronous Learning Networks 16(3): 9-20.