A Systematic Review On Disease Prediction In Big Data With Optimization And Map Reduce Framework

K.Manohari
Research Scholar
Department of Computer Science
Theivanai Ammal College for Women (Autonomous), Villupuram, Tamil Nadu, India

Dr.S.Manimekalai
Head,Department of Computer Science
Theivanai Ammal College for Women (Autonomous), Villupuram, Tamil Nadu, India

ABSTRACT

The vast amount of health data continues to expand every second, making discovering some sort of valuable knowledge harder and very challenging. Traditional way of providing data, particularly in health sector, has recently been changed by big data into useful insights. It offers a wide variety of preventive advantages for initial diagnosis of critical illnesses and provides quality health care to right patient. It has created resources for rapid collection, management, review and assimilation of large amounts of disparate, organized and unstructured vital data generated by various storage systems of health information. In existing health data analytics systems, though, there are many problems to be tackled that provide strategic methods such as critical data collection, aggregation, operation, interpretation, simulation, and sharing. Due to lack of detailed analysis in existing research works, this paper examines available methods in terms of 3 categories such as disease prediction, optimization algorithm and Hadoop map reduce framework in big data with their comparative analysis. This systematic review will provide a clear idea for researchers about various methods for particular disease prediction in big data.

Keywords–*Big data, disease prediction, optimization, map reduce, Diabetics*

INTRODUCTION

From compliance, patient related data and keeping, healthcare industry produces large amount of data. This data is digitized mandatory in today's digital world. To answer new challenges, it is significant to generate large amount of data, to improve healthcare quality by reducing costs [1]. Every day, government produces data petabytes.To perform real-time analysis on enormous data set, it needs technology [2]. To provide value-added services to citizens, will help government. By understanding data patterns as well as relationships among them with help of ML methods, big data analytics helps in finding valuable decisions [3].

Doung Laney is characterized by velocity, variety and volume called 3Vs called big data notion [4].Big data is defined as huge data collection with wide range of types. Using conventional database management system it is very hard to process.Big data contains data sets such as high speed and diversity and large volume that needs new style processing [5]. Massive data is referred to as big data while analyzing, visualizing and capturing data with current technologies are overwhelming. Due to advancement of healthcare technologies, big data plays a significant role in current digital era [6]. Big data sources are concerned in various sectors and healthcare industries are well known for diversity and their volume. Through big data impact, healthcare domain gained its effect. Over past couple of years, healthcare industries produce enormous amount of healthcare data. In terms of characteristics, healthcare data are same as big data, then it is called healthcare big data. Healthcare data incorporates EMRs (Electronic Medical Records) like physician notes, biometric data, patients' medical history, clinical reports and other medical data related to health.Both these data together lead to large-scale medical data together with efficient medical applications for large-scale data rely solely on the infrastructure underpinning it. It also offers an idea for big data research on health systems [7].

By developing the need for big data in medical use, this paper provides an overview of big data analytics in healthcare.

- Providing patient-centric services: Providing patients a quick relief by the evidence-based treatment that identifies infections early in the world based on available health results. Reducing medicine doses to prevent a side-effect. This decreases readmission rates and reduces patients' expenses[8].
- Detecting spreading diseases earlier: Prediction of viral diseases until they propagate based on live analysis. The social records of the patients who suffer from a disorder can be studied in a specific area to identify this fact. This helps health workers to warn victims by taking preventive steps.
- Monitoring the hospital's quality: Track the establishment by the Indian Medical Council of the hospitals according to guidelines. This daily check lets the government take appropriate steps to combat disqualification in hospitals [9].
- Improving the treatment methods: The impact of pharmaceutical products continuously and dependent on research can be continuously tracked for quicker relief by personalized patient therapy. Monitoring of critical patient signs for proactive patient treatment. The study of the patients with the same effects by evaluating the results allows a specialist to supply new patients with appropriate medicines.

Survey on disease prediction in big data

Joo et al.,(2020) Creation and evaluation, through different metrics like receiver operational characteristic curves, precision and recall curves, accuracy, specificity and F1 score, of different ML prediction methods that use logistic regression, DNN, LightGBM and random forests.The author has analyzed the cohort of Big Data for the Korean National Health Sample and proposed different models of ML prediction to estimate cardiovascular disease (CVD) risk 2 years and 10 years.Both suggested ML models were substantially higher than the baseline approach extracted from the guidance when we used the past drug knowledge as input functions.Past medication data is not included as main drawback [10].

Santosh et al.,(2020)suggest a new method for scaling malaria cases in chosen geographical sites to be projected. In Telangana, India, malaria abundances were predicted using satellite data as well as clinical data, along with the Long Short-Term Memory (LSTM) classification. The model recommended for selected regions in State established a seasonal pattern of 12 months. The findings revealed that the vast volume of data with low latency as well as scalability can be processed and analyzed[11].

Wang et al.,(2019) Present a multi-disease risk modeling process Neural methodology (NE) for the systemic assessment of possible disease risks to patients based on their medical demographic data.The research aggregates medical diagnoses based on the International Disease Classification (ICD) to different levels to estimate the needs of various players. Two separate medical datasets, including 710 5 patients with 18, 893 and 4170 patients with 124 visits respectively, have been used to verify the suggested approach [12].

Usama et al.,(2018) Propose a modern multimodal disease risk assessment focused on the RCNN using intra-layer recurrent, convolutional layer data from hospital to be a bidirectional RNN. Feed forwards and repeated inputs of previous unit as well as neighborhood are delivered to each neuron in the convolution layer.The context-capture area rises, not only step by step but also the extraction of fine-grain features.The downside of this work is that it would take more time to find[13].

Chen et al.,(2017)developed new CNN-MDRP (Convolutional Neural Network-based Multimodal Disease Risk Prediction) algorithm that combines hospital structured and unstructured data. The goal is efficient anticipation and overcoming of the challenge of insufficient data using a latent factors model to recreate lost data for chronic disease epidemic in populations with recurrent disease.The drawback is that in different areas, mainly because of the various climatic and living practices in the world, there is a significant variation between diseases[14].

Survey on optimization algorithm in big data

AbdElaziz et al.,(2020) built multi-target Big Data optimization approach based on a method and differential method of hybrid salp swarm.The purpose of the DE algorithm is to improve the capability of running Salp swarm method because operators of DE method are utilized as LSO.Suggested approach usually comprises three steps. The first stage is to create the population and to configure the directory.In the second phase solutions are updated using a combination salp swarm method and differential progress algorithm and in final phase solution is calculated and attain is updated[15].

Banchhor et al.,(2020) The Cuckoo-Grey Wolf Correlative Naive Bayes classifier and MapReduce Model has developed a method for classification of big data (CGCNB-MRM).The CG-CNB is designed to change CNB Classifier using a newly developed optimization method. The classifier is also new (CGWO). CGWO is designed to optimize the CNB model through the optimum collection of model parameters by successful integration of the Cuckoo Search (CS) method into GWO.Finally, in conjunction with the Likelihood Index tab and post-probability data the suggested CGCNB-MRM method classifies the data samples for each sample [16].

Ding et al.,(2019) In particular, LDA-based diabetically complicated subject mining is based on similarities between textual medical reports after data preprocessing.Propose solution to prediction of diabetic complication, based on a Latent Dirichlet Assignment (LDA) method strengthened by similarity. Finally, by solving a multi-mark classification issue with SVMs we are building a prediction model [17].

Babu et al.,(2018)predict various disease forms using the grey Wolf Optimization and RNN (GWO+RNN) auto-encoder.Using GWO, the characteristics are chosen and diseases are predicted using RNN system.After properties are passed to the RNN classifier, GWO method initially greatly avoids irrelevant as well as redundant attributes.Experimental outcome showed that GWO+RNN algorithm efficiency was higher than current methods such as Group Search Optimizer and Fuzzy Min-Max NN (GFMMNN)[18].

Nalluri et al.,(2017)developed hybrid technique to diagnose ailments using two classifier techniques, namely SVM and MLP techniques, to optimize individual classifier parameters.To optimize specifications of above classifiers, we use three recent evolutionary algorithms, leading to six alternative diagnostic systems for hybrid diseases, also known as Hybrid Intelligent Systems (HISs). On 11 Benchmark datasets, the model proposed is tested and the findings obtained indicate that hybrid diagnostic systems perform better in terms of precision, specificity and sensitivity of disease predictions [19].

Survey on Hadoop map-reduce framework

Choi et al.,(2019) propose a big-data health-related information process using Association Mining tools Hadoop's Map Reduce.The approach proposed provides effective knowledge management services via the collection and processing of heterogeneous health information through WebBot and common model of data.Hadoop is a patented means of managing distributed massive data efficiently. It is paradigm of information management incorporating distributed MapReduce-based processing and technique of discovering connections based on mining.Data in MapReduce was drawn from the nomenclature of chronic diseases from broad health data.Big data is divided into multiple blocks of size and map tasks generated [20].

Nair et al.,(2018)goal are to build a real-time remote health status detection framework based on Apache Spark's open-source Big Data processing engine deployed in cloud, which focuses on implementing Big Data streaming ML concept.Machine learning algorithm to anticipate health status of user.Research provides for consumers to convenience their areas with trouble-free, real-time digital health observation, without unnecessary costs [21].

Ramsingh et al.,(2018)powerful MapReduce-based NBC-TFIDF method is proposed to mine the feeling of individuals. NBC-TFIDF is the most successful MBS-Naive Bayes classifier algorithm.To distinguish data based on polarity score of every sentence in social media data.Using emotion corpus, polarity score is measured and diabetic corpus is generated utilizing food glycemic index and index of physical activity. This research uses social networking data to examine the association between food preferences. From the studies, it is clear that those who would be affected by diabetes in general are the younger generations (users of social networking sites) [22].

Dhanalakshmi et al.,(2017) suggested a new user recommendation framework has been developed to discourage the use of disease form, gene entities and user navigation habits for pranked biomedical papers.This recommendation method used the extremely applicable disease records on online PubMed repository for extraction of dynamic username recognition, dynamic user identification and documents ranking methods.The true positive rates and run-time of standard static models like Bayesian and Fuzzy have been compared to testing the efficiency of the proposed model. Experimental studies suggest that the model's efficiency is greater than typical ones [23].

Binu et al.,(2017) To provide cumulative knowledge, the proposed HADOOP method would integrate the illness and its symptom information and analyze it. After evaluating the resulting algorithm, a simple conclusion could be generated and sorting could be produced.As it wisely displays the data community, the method will be useful for a better view of the disease and its propagation rate, such that adequate care could be provided in due course [24].

Rodger et al.,(2015) Used the Hybrid Hadoop Hive patient computer processing software to orchestrate the processing of the database across separate boats, marshal the dispersed servers, execute multiple tasks simultaneously; handle all interactions and transitions between systems and maintain redundancy and defect tolerance. They have used the Apache hive to classify traumatic brain injury (TBI) and other accidents as an architecture built on the top of Hadoop for data summary, question or analysis. Finally, for data interpretation, a proposed method of Misdiagnosis Minimization Strategy was used [25].

COMPARATIVE ANALYSIS

This section contrasts disease prediction in large data by selecting parameters like precision, accuracy, recall, f-measure, and optimization algorithms in large data about parameters such as accuracy, sensitivity, speed convergence and Hadoop map by selecting the parameters, reducing framework methods of large data. These are all very strong, high, moderate, low and very low relative to the methods mentioned in Tables 1, 2 and 3.

Table1:Comparison of disease prediction methods in big data

Author/Year	Method	Accuracy	Precision	Recall	F-Measure
Joo et al.,(2020)	Machine Learning (ML)-based prediction	high	low		moderate
Santosh et al.,(2020)	Long Short-Term Memory (LSTM)	low	moderate	Very high	Very low
Wang et al.,(2019)	Neural Approach (NA)	low	moderate	Very low	low
Chen et al.,(2017)	Convolutional Neural Network-based Multimodal Disease Risk Prediction (CNN-MDRP) algorithm	Very low	Very high	moderate	low
Usama et al.,(2018)	Recurrent Convolutional Neural Network(RCNN)	high	moderate	high	low

Table 2: Comparison of optimization methods in big data

Author/year	**Method**	**Accuracy**	**Convergence speed**	**Sensitivity**	**Specificity**
Banchhor et al.,(2020)	Cuckoo-Grey wolf based Correlative Naive Bayes classifier and MapReduceModel (CGCNB-MRM).	moderate	Very low	low	moderate
AbdElaziz et al.,(2020)	hybrid salp swarm algorithm and the differential evolution algorithm	low	very low	moderate	Very high
Ding et al.,(2019)	similarity-enhanced Latent DirichletAllocation (seLDA) model	Very high	Very low	low	high
Nalluri et al.,(2017)	Support Vector Machine (SVM) and Multilayer Perceptron (MLP) technique.	Very low	Very high	moderate	low
Babu et al.,(2018)	Grey Wolf Optimization and autoencoder based Recurrent Neural Network (GWO+RNN)	moderate	high	Very high	low

Table 3: Comparison of map-reduce framework in big data

Author/year	**Method**	**Computational time**	**Complexity**	**Reduction ratio**
Choi et al.,(2019)	Hadoop’sMapReduce software for association mining	moderate	Very low	high
Nair et al.,(2018)	Apache Spark	low	Very high	moderate
Dhanalakshmi et al.,(2017)	novel user recommendation system	high	low	Very high
Ramsingh et al.,(2018)	MapReduce-Based Hybrid NBC-TFIDF (Naive Bayes Classifier -Term Frequency Inverse Document Frequency) algorithm	low	moderate	Very low
Binu et al.,(2017)	HADOOP system	Very low	Very low	moderate
Rodger et al.,(2015)	Patient Informatics Processing Software Hybrid Hadoop Hive	moderate	Very high	low

CONCLUSION

Previous literature relies on this systematic analysis to analyze big data in healthcare focused on established keywords and research aspects. This research provides the understanding of the different models available and multiple optimizations, a study carried out between 2015 and 2020 reduces the approaches used in big data. The compared parameters are accuracy, precision, recall, f-measure, responsive characteristics, specificity, convergence speed, time, complexity and reduction ratio. This survey will provide researchers with a valuable framework for future studies to consider the broader meaning and uses of big data in healthcare. Future work is to classify key patterns and characteristics of the patient's medical data by integrating optimization-based machine learning approaches with Big Data Analytics to forecast diabetics in teenagers before they are used by clinicians. The other aim is to reduce the data sets and enhance the prediction model accuracy.

REFERENCES

[1] R. Das, I. Turkoglu and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles",*Expert Systems with Applications*, vol.36, no.4, pp.7675-7680, 2009.

[2] N. G. Hedeshi and M. S. Abadeh, "Coronary artery disease detection using a fuzzy-boosting PSO approach",*Computational Intelligence and Neuroscience*, 2014.

[3] H. Yan, Y. Jiang, J. Zheng, C. Peng and Q. Li, "A multilayer perceptron-based medical decision support system for heart disease diagnosis",*Expert Systems with Applications*, vol.30, no.2, pp.272-281, 2006.

[4] M. Viceconti, P. Hunter and R. Hose, "Big data, big knowledge: big data for personalized healthcare",*IEEE Journal of Biomedical and Health Informatics*, vol.19, no.4, pp.1209-1215, 2015.

[5] Y. Zhang, L. Zhang, E. Oki, N. V. Chawla and A. Kos, "IEEE Access special section editorial: big data analytics for smart and connected health",*IEEE Access*, vol.4, pp.9906-9909, 2016.

[6] S. R. Sukumar, R. Natarajan and R. K. Ferrell, "Quality of big data in health care",*International Journal of Health Care Quality Assurance*, vol.28, no.6, pp.621-634, 2015

[7] T. R. Hoens, M. Blanton, A. Steele and N. V. Chawla, "Reliable medical recommendation systems with patient privacy",*ACM Transactions on Intelligent Systems and Technology*, vol.4, no.4, pp.1-31, 2013.

[8] Collobert R, Weston J, Karlen M, Kavukcuoglu K and Kuksa P, "Natural language processing (almost) from scratch", *Journal of Machine Learning Research*, vol.12, pp.2493-2537, 2011.

[9] Lee C,LinW,ChenY,KuoB,"Gene selection and sample classification on microarray databased on adaptive genetic algorithm/k-nearest neighbor method",*Expert Systems Applications*, vol.38, 2011.

[10] Joo G, Song Y, Im H and Park J, "Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea)", *IEEE Access*, vol.8, pp.157643-157653, 2020.

[11] Santosh T, Ramesh D and Reddy D,"LSTM based prediction of malaria abundances using big data", *Computers in Biology and Medicine*, vol.124, 2020.

[12] Wang T, Tian Y and Qiu R. G,"Long Short-Term Memory Recurrent Neural Networks for Multiple Diseases Risk Prediction by Leveraging Longitudinal Medical Records", *IEEE Journal of Biomedical and Health Informatics*, 2019.

[13] Usama M, Ahmad B, Wan J, Hossain M. S, Alhamid M. F and Hossain M. A, "Deep feature learning for disease risk assessment based on convolutional neural network with intra-layer recurrent connection by using hospital big data", *IEEE Access*, vol.6, pp.67927-67939, 2018.

[14] Chen M, Hao Y, Hwang K, Wang L and Wang L, "Disease prediction by machine learning over big data from healthcare communities", *IEEE Access*, vol.5, pp.8869-8879, 2017.

[15] AbdElaziz M, Li L, Jayasena K. N and Xiong S,"Multiobjective big data optimization based on a hybrid salp swarm algorithm and differential evolution", *Applied Mathematical Modelling*, vol.80, pp.929-943, 2020.

[16] Banchhor C and Srinivasu N,"Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification", *Data & Knowledge Engineering*, vol.127, pp.101-788, 2020.

[17] Ding S, Li Z, Liu X, Huang H and Yang S,"Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model", *Information Sciences*, vol.499, pp.12-24, 2019.

[18] Babu S. B, Suneetha A, Babu G. C, Kumar Y. J. N and Karuna G,"Medical disease prediction using grey wolf optimization and auto encoder based recurrent neural network", *Periodicals of Engineering and Natural Sciences*, vol.6, no.1, pp.229-240, 2018.

[19] Nalluri M. R and Roy D. S,"Hybrid disease diagnosis using multiobjective optimization with evolutionary parameter optimization", *Journal of healthcare engineering*, 2017.

[20] Choi S. Y and Chung K,"Knowledge process of health big data using MapReduce-based associative mining", *Personal and Ubiquitous Computing*, pp.1-11, 2019.

[21] Nair L. R, Shetty S. D and Shetty S. D,"Applying spark based machine learning model on streaming big data for health status prediction", *Computers & Electrical Engineering*, vol.65, pp.393-399, 2018.

[22] Ramsingh J and Bhuvaneswari V,"An efficient Map Reduce-Based Hybrid NBC-TFIDF algorithm to mine the public sentiment on diabetes mellitus–A big data approach", *Journal of King Saud University-Computer and Information Sciences*, 2018.

[23] Dhanalakshmi P, Ramani K and Reddy B. E,"An improved rank based disease prediction using web navigation patterns on bio-medical databases", *Future Computing and Informatics Journal*, vol.2, no.2, pp.133-147, 2017.

[24] Binu P. K, Akhil V and Mohan V,"Smart and secure IOT based child behaviour and health monitoring system using hadoop", *International Journal of Engineering Research & Technology(IJERT)*, pp.418-423, 2017.

[25] Rodger J. A,"Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive", *Informatics in Medicine Unlocked*, vol.1, pp.17-26, 2015.