Healthcare Analytics Using Big Data for Evaluation and Extreme Machine Learning Based on MapReduce

R. Ramesh[1] *and V. Selvam*[2*]

Abstract

The present study focuses on Extreme Machine Learning Method (ELM), which may be used with Support Vector Machines that are optimised by Cuckoo Search to produce a method for identifying disease risk (CS-SVM). It also considers the accuracy and scalability of big data models, which considerably increase the processing power of the proposed method and produce better outcomes in terms of performance metrics like veracity and efficiency. In terms of additional performance metrics like Precision, Recall, and Average Area Under Curve, the suggested method is also compared to comparable cutting-edge methods.

Keywords : **Big data, Extreme Machine Learning**

I. INTRODUCTION

A range of datasets that are scattered widely and are yet retained are referred to as "big data." In big data healthcare, Electronic Health Records (EHR) are used to manage clinical data sets of each patient's information. Clinical data is kept in a number of various formats and is unstructured to a degree of over 80% [1]. Processing huge information creates challenges and expectations for data storage and analysis in terms of efficiency and scalability. Big data is used by the Hadoop MapReduce architecture to quickly store and process all types of data. It is adaptable to the systems and fault-tolerant. Also, a Machine Learning algorithm handles the prediction of data sets. The size of data is growing in step with the rapid growth of technology. People currently live in a world that is driven by data. Massive storage volumes in the petabyte range are organised into data sets. The medical sector uses a variety of data in structured, semi-structured, and unstructured forms while examining big data. Raw data examples include complicated reports, a patient's medical history, and the results of electronics tests. These medical reports contain both structured and unstructured data. Structured data can easily be incorporated into risk prediction models [2]. However, due to the discontinuity, complexity, multidimensionality, and noise of this data, unstructured data formats are hiding a tonne of crucial information. In the healthcare sector, Electronic Health Records (EHS), which contain patient disease records and enhance clinical decision-making are used. The information from clinical healthcare can be used for predictive analysis. However, unstructured data formats are concealing a tonne of important information because of the discontinuity, complexity, multidimensionality, and noise in this data. Electronic Health records (EHS), which store patient disease records and improve clinical decision-making, are employed in the healthcare industry. Predictive analysis can be done using data from clinical healthcare. By using diverse analytical techniques, the healthcare sector offers immense potential to cure diseases more successfully. Machine

Manuscript Received : January 6, 2023 ; Revised : January 16, 2023 ; Accepted : January 19, 2023. Date of Publication : February 5, 2023.

R. Ramesh[1], *Associate Professor*, KPR College of Arts Science and Research, Avinashi Road, Coimbatore - 641 407. Email : Proframeshr@gmail.com ; ORCID iD : https://orcid.org/0000-0001-8017-8551

V. Selvam[2*], *Assistant Professor*, KPR College of Arts Science and Research, Avinashi Road, Coimbatore - 641 407. Email : vselvamphd@gmail.com ; ORCID iD : https://orcid.org/0000-0003-3617-2878

DOI : https://doi.org/10.17010/ijcs/2023/v8/i1/172682

Learning (ML) approaches such as Naïve Bayesian (NB), Decision Tree (DT), KNN, and Neural Network (NN) are used to manage the structured data. The problem with healthcare data is the processing of unstructured data. Massive data sets have proven to be challenging to train for unsupervised learning in a number of researches on unstructured data processing [3].

The basis for data distribution is provided by the Hadoop MapReduce framework. The master/slave architecture serves as its foundation. MapReduce offers scalable computations in a single iteration, which is one of its benefits. Some algorithms require cutting-edge data about one node's relationship to another at the moment of computation. Probabilistic Graphical Model (PGM) is a well-researched machine learning framework for examining various types of issue structures. A dataset used to predict the number of patients with the disease is used to assess the performance of the proposed strategy. Because it is free and open source, offers HDFS (Hadoop Distributed File System), which enables distributed storage, and is fault tolerant. The Apache Hadoop framework is widely utilised. To swiftly process massive volumes of data, *MapReduce*, a popular Hadoop programming framework can be employed. The training data and the testing data make up each half of the dataset in this instance. The system may respond intelligently to input thanks to the application of ML algorithms, producing significant data that is utilised to generate common reports in the processing layer. The ML techniques are mainly utilised to diagnose the condition. These machine learning techniques aid in making better decisions regarding treatment strategies and in offering superior healthcare services. There is a lot of information in clinical databases, including information on the patients, images related to radiographs, medical records, and other sensor data. It goes without saying that this information overflow will increase database size and complexity. When considering their uses in clinical healthcare research sectors, distributed systems like Hadoop and MapReduce are more favourable due to their vast storage capacity and capacity to process enormous amounts of data. Extreme Machine Learning is used to tackle issues in a wide range of therapeutic disciplines because of its capacity to extract valuable information from learned data.

II. ANALYTICS WITH LARGE-SCALE HEALTHCARE DATA

The majority of the time, patient-related data from the healthcare sector is made up of countless enormous and complex facts. The number and complexity of the data have grown even more due to current developments in the medical field, which also give sensor data collected by several electronic and mobile devices. By gathering clinical data from each patient at several hospitals, more information about the research is being conducted in a specific manner. Last but not the least, records of information including clinical images, CT scan and MRI scan results, lab records, surgical records, and insurance information are periodically updated in the clinical databases. To produce the intended results, the records used in hospitals—including clinical pharmacy images, electronic data like X-ray images, photos from MRI scans, post-surgery reports, information about medicine, and other general data are used. As a result, the size of the

TABLE I.

BIG DATA IN HEALTH CARE SYSTEMS

DATA CATEGORY	LISTS	DESCRIPTION
STRUCTURED DATA	Details of patient	Age, sex, height, weight, etc. of the patient.
	Habitual details	Records of genetic information, smoking, drinking, and other behaviours.
	Examination reports	Reports from blood tests, BP tests, etc.
	Predicted disease reports	Disease histories, including diabetes, blood pressure, etc.
	Pill prescription details	Kind of medication, then patient information.
UNSTRUCTURED TEXT DATA	Patients readme illness	Medical background.
	Doctors Details	Details of the doctor's interrogation.
	Medical appointment details	How often the patient has been enrolled for a check-up.

healthcare database is dramatically growing. Data that is utilized for training has the most Electronic Health Record (EHR) data (nearly around 80% from the data set). The learning machine network's input features are regulated and prediction activities are carried out using training data sets. The test data that only test the final output are used to confirm the algorithm's true predictive capabilities. 20% of the clinical data set is contained in the test data.

III. RISK FACTOR PREDICTION

The risk factor is typically found using Machine Learning and deep learning methods. The primary emphasis of this study is on accuracy measurements. We combine the CS-SVM method with the ELM algorithm to increase accuracy and optimization [4]. The machine learning methods Naïve Bayes (NB), CS-SVM, and Decision Tree (DT) are used to estimate the risk of fatal disease using structured data. We improved the ELM to create an Optimized Extreme Learning Algorithm for Unstructured Data in order to predict fatal disease based on the training data sets. Classification algorithms like CS-SVM, DT, and NB [6–7] are utilised to provide the prediction for structured data. A data set of dengue fever patients is gathered from several hospitals in order to determine the risk factor for dengue fever from medical records. There are 820 patient records in the structured data set, and each record has a total of 20 label attributes. The target class's label attribute has nominal values. These figures are derived from a variety of data including test outcomes, X-rays, the patient's medical history, and other relevant details. The examination's data, including the heart rate, blood pressure, and outcomes of additional laboratory tests are used in the feature selection process.

The classification that is based on probability is the Naïve-Bayes classification. Calculating the likelihood in the attributes of the chosen features is necessary for this. The Gaussian distributive function and the probability formulae for discrete feature estimation and attribute estimation are used respectively. The patient data set is randomly split into a test data set and training data set with a ratio of 5:1. The MapReduce approach is used to carry out the categorization simultaneously.

IV. ALGORITHMS OF MACHINE LEANING

A. A More Effective Naïve Bayesian Classifier

It is a classification method [8] that uses a statement of unpredictability among prediction analysts and is based on the Bayes theorem. An individual feature inside a class is recognised by a Naïve Bayes classifier as being distinct from all other features. In order to produce predictions, an NB classifier normally considers each feature characteristic separately and will increase the likelihood. By allowing the depiction of interdependence among subsets of characteristics, the Augmented Naïve Bayesian classifier (INB) that has been described classifies the data sets.

From Fig. 1, The probability parameter P(a|b) is produced by the factors P(a), P(b), and P(b|a), where given a known predictor attribute, the posterior probability of the predictor class is P(a|b). P illustrates the prior probability of the class (c). P(b) is the prior probability of the predictor, and P(b|a) is the likelihood, which is the predictor's probability for a specific class.

The computation of posterior probability using Naïve Bayes is the prediction's outcome. The updated Naïve

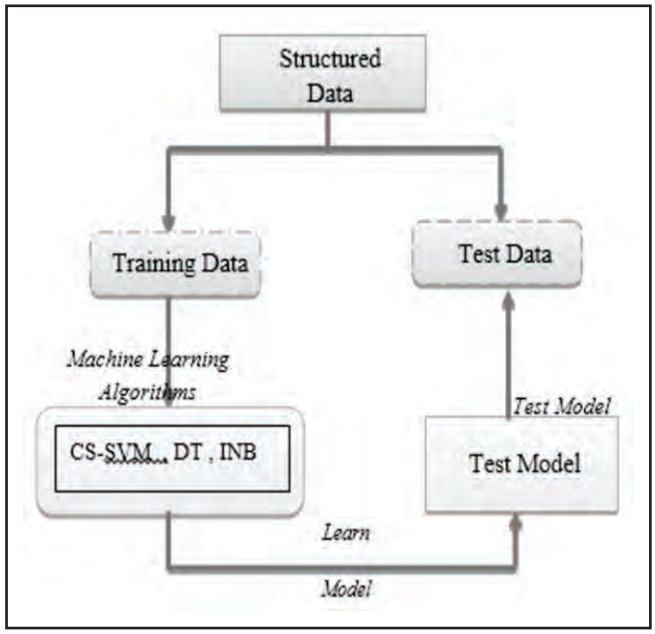

Fig. 1. Machine Learning Techniques for Structured Data Prediction

Bayesian model has good accuracy due to the independent values of the attributes in the healthcare data. This statistical model has a high level of accuracy. This strategy is based on the idea that each quality exists independently of the others. This classifier can function well in the medical field with or without preprocessing. The method works wonders on a variety of challenging issues.

V. DECISION TREE

There are three essential conditions for separating data in this supervised learning methodology [9]. The most crucial of the three conditions is knowledge acquisition. Entropy is defined in Equation 1 for a problem involving many classes

$$E(D) = \log 2Px \qquad (1)$$

where,

D Training data ;

Px Discretization value of test data .

When there is no more information than there is space, the entropy is determined. This action should be taken after a multi-stage decision-making process.

VI. CUCKOO SEARCH-SVM

A statistical learning method that maps the data into a high-dimensional space F using a non-linear mapping function (X). The categorization is improved in a high-dimensional space using linear regression. F(a) mapping function shown in Equation 2.

$$f(a) = \omega . \phi (A) + b = 0 \qquad (2)$$

where,

b Threshold parameter ;

ω Weight vector ;

A Distance vector.

The Cuckoo Search algorithm is a potent optimization method that conforms to three core concepts. Every cuckoo species is working on a special approach to increase the likelihood that its eggs will hatch [6]. There are three main rules in CS.

- Each cuckoo in the nest will randomly place an egg in a nest at any point throughout a predetermined period of time when she is expected to lay an egg.
- The next generation will receive only the best eggs.
- Using the host bird with a probability of Pa £, the eggs-laying birds are found after settling on a small selection of more significant hosts (0, 1).

With the use of the birds in the nested host, this algorithm has been refined to produce better outcomes. It is clear that the best search path and the location of the bird's nest are frequently updated by the CS searching algorithm.

This is shown in Equation (3) as

$$Levy = Xis + Xi\, s + 1 \qquad (3)$$

Where,

$Levy$ Best search path;

Xis Set of birds ;

Xi An instance of a set.

In order to achieve classification, training data sets are acquired, the CS-SVM is initialised with the probabilistic parameter Pa = 0.75, and a randomly generated nest N location is employed. Before determining the current best location, the training set is used for cross validation. A classification model is created based on the test set after the test set is gathered using SVM.

VII. ALGORITHM FOR CLASSIFYING DISEASE RISK BASED ON EXTREME LEARNING METHOD

In order to manage high data-speed with quick learning, outstanding performance, and low computer complexity, the Extreme Learning Method (ELM) [10] was developed. The data currently available in the health sector are largely unstructured when it comes to Extreme Machine Learning, even with the adoption of Learning Method (LM) and a data platform. With no requirement to retrain the data pattern, ELM learning is a feed-forward network that alters the sum of the input and output in any logical way [11]. Since it has the power to change the purpose of the underlying logic at any time, the difficulty of scalability is important for any business. For the

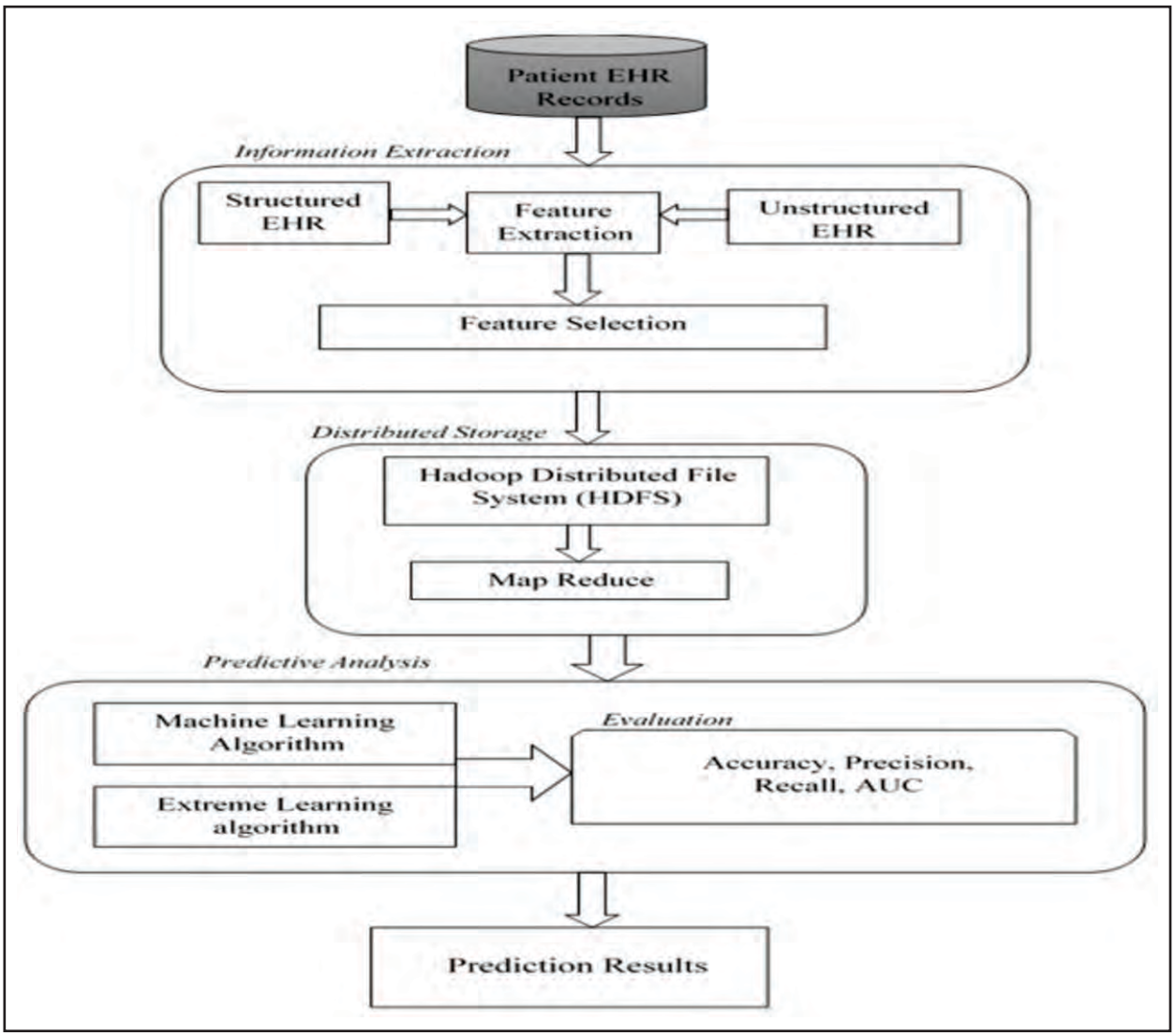

Fig. 2. Proposed Block Diagram's Overall Process

prediction of data, a scalable system is provided that led to improved performance with little time complexity [12]. A scalable learning system for supervised learning that takes into account unstructured data is developed by the suggested ELM. The outcome, which is undoubtedly unknown given the input, will be anticipated by the taught system. The first challenge while using Learning Method (LM) is the pre-processing of data, which includes operations like normalisation, missing value resolution, data transformation, feature extraction, and data training. The main goal of data analytics is to extract useful information from the data by identifying all potential links between the data.

Fig. 2 depicts the overall process of Extreme learning method (ELM). In addition to altering real-time training and prediction, other learning methodologies have an effect on each machine's performance by reducing the level of performance metrics attained. On data sets of a realistic size, our proposed system outperforms existing approaches in an excellent manner.

VIII. SELECTION FEATURE FROM UNSTRUCTURED ELECTRONIC HEALTH RECORD

Map feature in A. ELM

For generalised Single Hidden Layer Feed Forward Neural Networks, the output function of the ELM

network topology is calculated using the formula. $F(x)$, the output function is shown in Equation (4).

$$F(x) = (x) = h(x) \quad (4)$$

Where,

$h(x)$ Hidden layers output vector between the hidden layer and the output layer with $m > 1$ output nodes.

(x) Hidden layer.

$$h(x) = [h1(x),..hL(x)] \quad (5)$$

Where,

$h(x)$ Hidden layer's output vector ;

$h1(x)$ Hidden layer output;

$h1(x)$ is referred as the nonlinear ELM feature mapping classifier is taking risks into account and identifying potential risk factors.

IX. RESPONSE AND CONVERSATION

Since CS-SVM has less optimization restrictions and is also simpler to implement, the suggested traditional optimization procedure in ELM can be linearly extended to it. Each dataset was subjected to various categorization methods, such as Neural Network, Logistic Regression, Random Forest, Naïve Bayes, and Cuckoo Search-Support Vector Machine. The statistical measurements for various classification algorithms are covered in Table II. Predictive analytics may be measured using the True Positive (TP), False Positive (FP), Positive Prediction (PP), and accuracy factor metrics. 99% of the scheduled work will be completed. The extended learning machine and machine learning classifiers are applied to 2 Gigabyte, 6 Gigabyte, and 12 Gigabyte datasets as a test data set, and they are deployed in a high performance distributed network. Platforms and procedures for high performance computing are needed for this.

The distributed framework's speedup is shown in Fig. 3 with each cluster containing a sizable number of data sets, which allows the system to operate more quickly. Due to the increasing parallelization of the healthcare data sets, map reduction delivers a higher success rate as the size of the data sets increases.

Fig. 4 is based on cluster series that are run on various numbers of nodes and calculates the execution time for each test data set. The size of the input data collection has a nearly linear effect on the execution time. The suggested approach aids in shortening the system's execution time and increasing efficiency.

X. CONCLUSION

In really vast data sets, ELM is likely to persist. The focus of this research is on accurate disease prediction using both structured and unstructured data sources. The data storage and solutions would much rather provide a better solution than the conventional storage techniques. The algorithm and services can be improved to further the

TABLE II.

PERFORMANCE MEASUREMENTS FOR DIFFERENT ML ALGORITHMS

Classification Approach	Precision	Recall	Accuracy	AUC
Neural Network (NN)	85.2	88.1	83.8	91.9
Random Forest	67.3	94.3	67.3	88.1
SVM	85.8	86.7	82.4	92.4
Decision Tree(DT)	87.9	81.4	87.1	88.3
Logistic Regression	83.9	91.2	81.2	86.9
Naïve Bayes	91.2	95.4	88.6	95.5
Improved Naïve Bayesian	92.3	97.3	92.4	98.3
CS-SVM	93.6	92.5	93.5	97.9

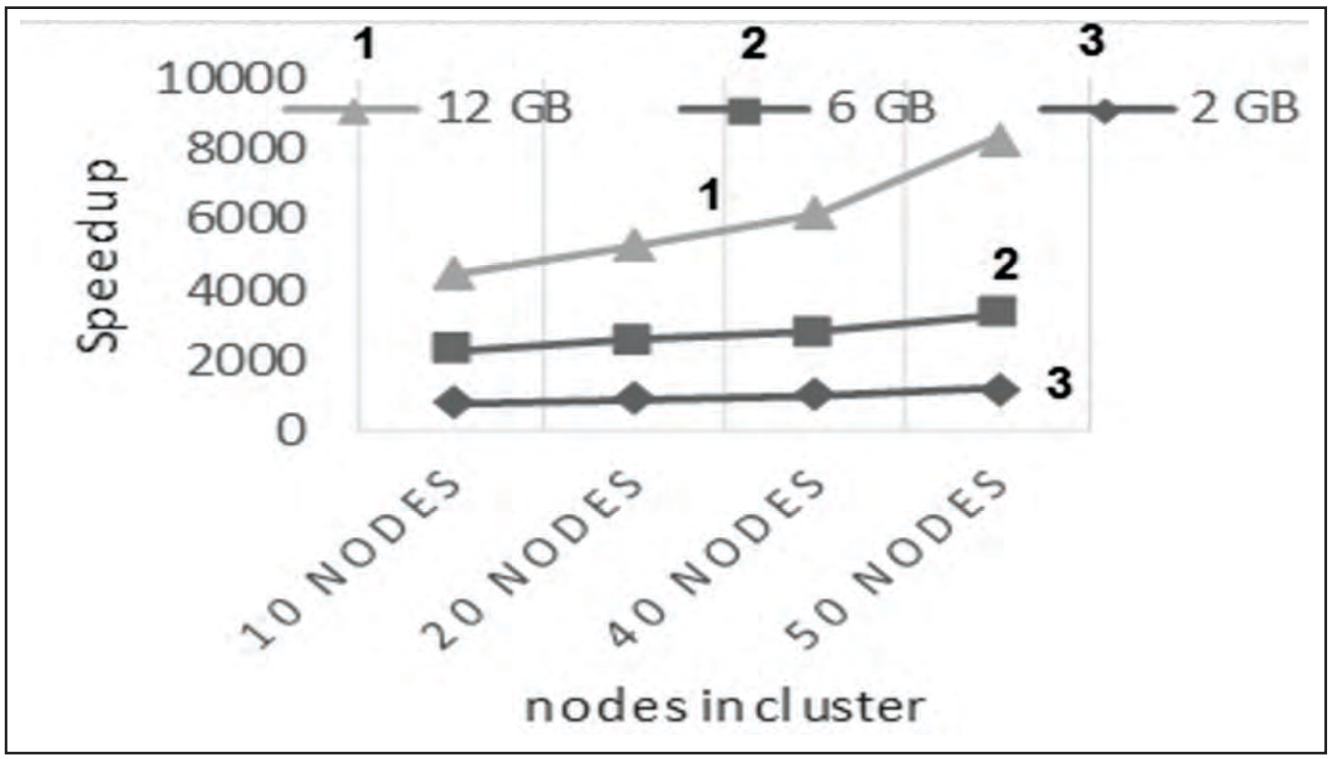

Fig. 3. Accelerations for Every Data Set

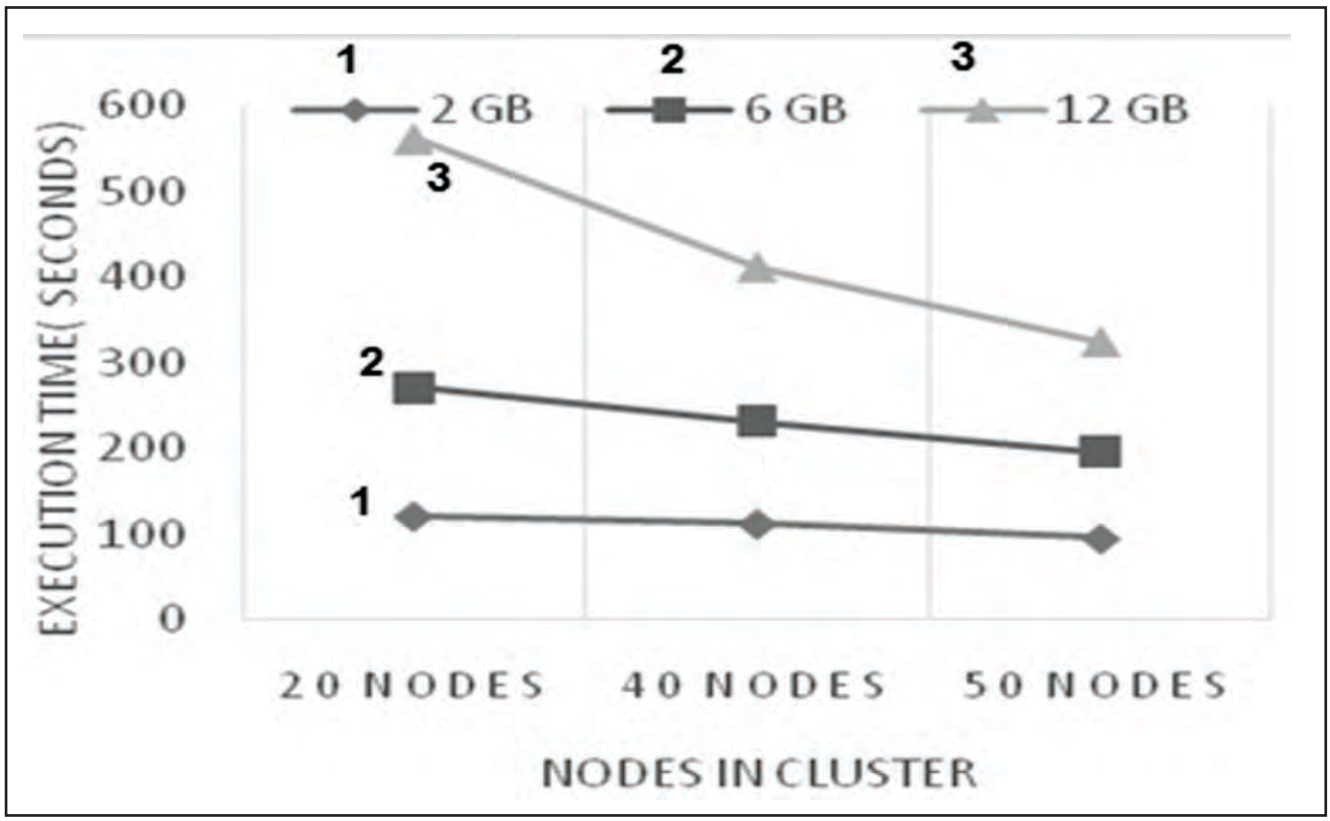

Fig. 4. Execution Time for Data Sets

research and produce better outcomes. Additionally, suggested work utilising MapReduce for unbalanced data is concentrated on achieving scalability and fault tolerance in distributed platforms. To more accurately assess the risk of dengue fever, these two forms of data combined yield an accuracy rating of 98.70%.

AUTHORS' CONTRIBUTION

Dr. R. Ramesh conceptualized the research, designed research methodology, and also prepared the draft transcript. Dr. V. Selvam worked on the literature review, experimental results, and revised the draft. Both the authors collectively finalized the article.

CONFLICT OF INTEREST

Dr. R. Ramesh and Dr. V. Selvam certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in the manuscript.

FUNDING ACKNOWLEDGEMENT

Dr. R. Ramesh and Dr. V. Selvam have not received any financial support for the research, authorship, and/or for the publication of the article.

REFERENCES

[1] F. Ö. Çatak, "Classification with Extreme Learning Machine and ensemble algorithms over randomly partitioned data," in *23rd Signal Process. Commun. Appl. Conf.*, 2015, pp. 228–231, doi: 10.1109/SIU.2015.7129801.

[2] J. -H. Zhai, S. -F. Zhang, M. -H. Wang, and Y. Li, "A three-stage method for classification of binary imbalanced big data," in *2020 Int. Conf. Mach. Learn. Cybernetics*, Adelaide, Australia, 2020, pp. 207–212, doi: 10.1109/ICMLC51923.2020.9469568.

[3] P. Nair and I. Kashyap, "Hybrid pre-processing technique for handling imbalanced data and detecting outliers for KNN classifier," in *2019 Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput.*, 2019, pp. 460–464, doi: 10.1109/COMITCon.2019.8862250.

[4] M. S. Hossain and G. Muhammad, "Healthcare Big Data Voice Pathology Assessment Framework," in *IEEE Access*, vol. 4, pp. 7806–7815, 2016, doi: 10.1109/ACCESS.2016.262631.

[5] H. Wen and F. Huang, "Personal loan fraud detection based on hybrid supervised and unsupervised learning," in *2020 5th IEEE Int. Conf. Big Data Analytics*, 2020, pp. 339–343, doi: 10.1109/ICBDA49040.2020.9101277.

[6] S. Y. Choi and K. Chung, "Knowledge process of health big data using MapReduce-based associative mining," *Personal Ubiquitious Comput.*, vol. 24, no. 5, pp. 571–581, 2020.

[7] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *J. Big Data*, 1, Article No. 2, 2014, doi: 10.1186/2196-1115-1-2.

[8] C. K.-S. Leung, R. K. MacKinnon, and F. Jiang, "Finding efficiencies in frequent pattern mining from big uncertain data," *World Wide Web*, vol. 20, no. 3, pp. 571–594, 2017.

[9] H. Guo, H. Liu, J. Y. Chen, and Y. Zeng, "Data mining and risk prediction based on apriori improved algorithm for lung cancer," *J. Sign Process Syst.*, vol. 93, pp. 795–809, 2021, doi: 10.1007/11265-021-01663-1.

[10] P. K. Mantha, A. Luckow, and S. Jha, "Pilot-MapReduce: An extensible and flexible MapReduce implementation for distributed data," in *Proc. 3rd Int. Workshop MapReduce Appl. Date*, Association for Computing Machinery, New York, NY, USA, pp. 17–24, 2012, doi: https://doi.org/10.1145/2287016.2287020.

[11] P. Deligiannis, L. Hans-Wolfgang, and E. Kouidi, "Improving the diagnosis of mild hypertrophic cardiomyopathy with MapReduce," in *Proc. 3rd Int. Workshop MapReduce Appl. Date (MapReduce '12)*, Association for Computing Machinery, New York, NY, USA, 2012, pp. 41–48, doi: 10.1145/2287016.2287025.

[12] Y. Wu, L. Zheng, B. Heilig, and G. R. Gao, "Design and evaluation of a novel dataflow based bigdata solution," in *Proc. 6th Int. Workshop Program. Models Appl. Multicores Manycores*, Association for Computing Machinery, New York, NY, USA, 2015, pp. 40–48, doi: 10.1145/2712386.2712397.

About the Authors

Dr. R. Ramesh received Doctoral Degree from Bharathiar University and Master of Computer Applications degree from Karunya University. His career span of 13 years includes various academic and administrative responsibilities in renowned institutions across the state. His research credentials include publications in reputed journals. Dr. R. Ramesh has specialized in Data mining, Data analytics, and Web development.

Dr. V. Selvam has completed B.Com., M.Com., and M.Phil. from Gobi Arts and Science College, Gobichettipalayam, and has completed Ph.D. from Periyar University (regular mode) in Commerce discipline under the aid of Senior Research Fellow, University Grants Commission, New Delhi. He has teaching experience of 8 years and research experience of 13 years in various institutions. He has specialized in Marketing and Banking. He has guided 6 M.Phil. Research Scholars and is also guiding 3 Ph.D. research scholars. He has published papers in reputed journals. He has presented more than 59 research papers in national and international seminars and conferences.