



# A Systematic Literature Review and Future Perspectives for Handling Big Data Analytics in COVID-19 Diagnosis

Nagamani Tenali<sup>1</sup> · Gatram Rama Mohan Babu<sup>2</sup>

Received: 5 October 2022 / Accepted: 23 February 2023 / Published online: 16 March 2023

© The Author(s), under exclusive licence to The Japanese Society for Artificial Intelligence and Springer Nature Japan KK, part of Springer Nature 2023

## Abstract

In today's digital world, information is growing along with the expansion of Internet usage worldwide. As a consequence, bulk of data is generated constantly which is known to be "Big Data". One of the most evolving technologies in twenty-first century is Big Data analytics, it is promising field for extracting knowledge from very large datasets and enhancing benefits while lowering costs. Due to the enormous success of big data analytics, the healthcare sector is increasingly shifting toward adopting these approaches to diagnose diseases. Due to the recent boom in medical big data and the development of computational methods, researchers and practitioners have gained the ability to mine and visualize medical big data on a larger scale. Thus, with the aid of integration of big data analytics in healthcare sectors, precise medical data analysis is now feasible with early sickness detection, health status monitoring, patient treatment, and community services is now achievable. With all these improvements, a deadly disease COVID is considered in this comprehensive review with the intention of offering remedies utilizing big data analytics. The use of big data applications is vital to managing pandemic conditions, such as predicting outbreaks of COVID-19 and identifying cases and patterns of spread of COVID-19. Research is still being done on leveraging big data analytics to forecast COVID-19. But precise and early identification of COVID disease is still lacking due to the volume of medical records like dissimilar medical imaging modalities. Meanwhile, Digital imaging has now become essential to COVID diagnosis, but the main challenge is the storage of massive volumes of data. Taking these limitations into account, a comprehensive analysis is presented in the systematic literature review (SLR) to provide a deeper understanding of big data in the field of COVID-19.

**Keywords** Healthcare · Machine learning · Big data analytics · Deep learning · COVID-19 · Decision making

---

✉ Nagamani Tenali  
tenalinagamani@gmail.com

Extended author information available on the last page of the article

## 1 Introduction

Due to information technology advances, the modern society utilizes an enormous amount of data, making big data analytics an essential data management tool in every sector. In the twenty-first century, big data is ingrained in every aspect of contemporary life [2]. A key goal of big data technology is to predict future trends using observed patterns [8]. In order to provide large datasets, a considerable amount of data must be collected and advanced monitoring and analysis tools must be used. Conventional data processing systems have difficulty in storing, managing, or analysing big data collections [1]. To fix these issues, big data [3] come up with different characteristics like veracity, velocity, variety, volume, validity, variability, venue, vocabulary, vagueness, value [4]. Typically, Big data analytics is the methodology for analysing massive volumes of data to uncover pertinent information using cutting-edge techniques [11]. Massive amounts of data are now widely obtainable, setting new countless opportunities for heterogeneous datasets, facilitating better medical decisions, and enhanced efficiency in the healthcare system [12].

Academics may examine this vast amount of data using healthcare data analytics, spot patterns, and trends in the data, and propose a solution to improve healthcare, therefore reducing costs, modernizing access to healthcare, and saving irreplaceable human lives [13]. Big data technologies serve several sides of healthcare, including patient monitoring, clinical decision support and healthcare management [14]. Through the discovery of data correlations and the understanding of patterns and trends, big data technique offers a great ability to improve healthcare [7], and streamline the health care system's financial burden [15]. Clinicians, Epidemiologists, and health experts have a great potential to use big data to make choices based on the best evidence available, enhancing patient care [16]. Biomedical researchers must correctly comprehend and use big data, which is not just a current reality but also a requirement in the search for new information [5]. In order to effectually battle the COVID-19 epidemic in the current digital era, clear criteria for effective data collection and analysis on a global basis are crucial [17].

The COVID-19 outbreak has killed a huge number of people while wreaking havoc on society, the economy, and the health of the whole world. To effectively control an epidemic, it is important to understand its characteristics and behaviour [18]. This knowledge may be achieved by collecting and analysing the pertinent big data. Considering the vast quantity of COVID-19 [9] data that is currently accessible from several sources [19], it is vital to assess the roles played by big data analysis in preventing the COVID-19 extent as well as to outline the key challenges and potential paths for COVID-19 data studies going forward. Additionally, the COVID-19 [6] will provide improved treatments that are both economical and well-received by patients by combining big data with image processing techniques [20]. Therefore, the analysis of COVID-19 requires a framework based on existing applications and studies for providing early decision-making system [10]. Following is a list of contributions discussed in this review:

- Big data analytics in cancer disease-based systematic literature review is offered, acting as a road map for experts in the area to spot and deal with problems caused by new developments.
- A comprehensive analysis of the issues and challenges posed by deep learning-based healthcare big data analytics is given, along with a look ahead.
- A summary of the potential uses for deep learning methods in the context of evidence-based big data analytics is given in line with the discussion of open research challenges.
- This study enhances understanding and directs academics and experts toward recognising current developments and future directions in big data analytics by utilising deep learning techniques.
- Finally, the presented work provides the possible research guidelines to the upcoming research works related to COVID epidemic.

The following is how the manuscript is structured: Sect. 2 discusses about the background knowledge in context of big data analytics and covid-19 disease diagnosis along with a deep comprehensive study. Section 3 illustrates the analysis part and the further section explains about the future implications. The final Sect. 4 elaborates the conclusion of the study.

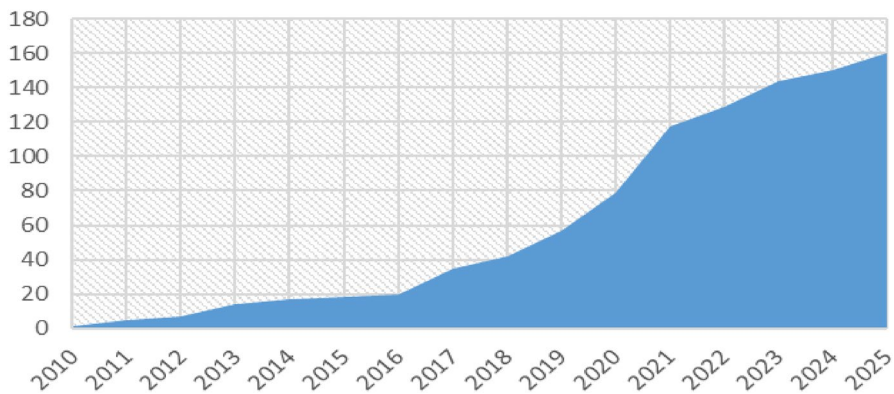
## 2 Background and Systematic Literature Review

The idea of “Big Data” has emerged due to the abundance of data usage in everyday life. The complexity of big data has a noteworthy impact on how well typical warehouses can collect, manage, monitor, and analyse it. Later on, big data, as defined by IBM’s big data analytics division, is a phrase used to describe datasets that are larger than those found in conventional databases. The dataset is created at enormous sizes and has significant variety, velocity, and volume. Big data aids analysts, researchers, and businesses in the decision-making process by applying a range of methodologies, including statistics, predictive analytics, machine learning, data mining, deep learning analytics, and text analytics. It is possible to use big data analytics in a wide variety of ways like Managing financial crises, medical research, Education, Banking, Natural language processing, and Data administration. Owing to the compensations of big data analytics, academia currently applies big data analytics specially to address the intractable issue like COVID disease identification. The comprehensive review of big data related to healthcare domain, big data characteristics, and issues associated with processing contemporary data are covered in the parts that follow (Fig. 1).

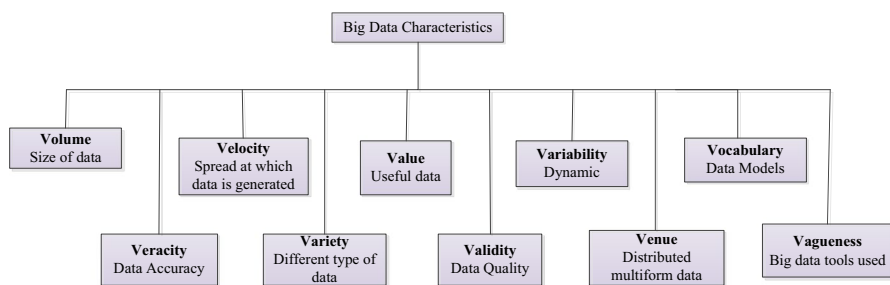
### 2.1 An Overview of Big Data: Definitions and Characteristics

Over the past few decades data has grown incredibly at an unforeseen rate. The volume of data is predicted to expand, as illustrated in Fig. 2, from a few zettabytes in 2010 to 163 zettabytes in 2025, according to the survey taken from International

## Volume of data in zettabyte



**Fig. 1** Volume of data transferred from 2010 and 2025 by IDC over decades



**Fig. 2** Characteristics of big data

Data Corporation (IDC). As a result, the capacity of data storage has expanded from megabytes to exabytes, and it is anticipated that it will approach zettabytes yearly in the coming years. As in prior, relational databases with rows and columns and ordered formats were used to store and display data, which was often obtained from internal operations. Many scholars now claim that the overwhelming of this data is unstructured, and handling it requires the use of non-relational (NoSQL) databases [37, 39]. This data may be categorized as machine-to-machine data, web human-generated data, web social media data, and biometric data and transaction data.

Because there are so many uses, academics have added additional elements to big data characteristics. Actually, 4Vs, 5Vs, 6Vs, 7Vs, and 10Vs have been added to the initial 3Vs concept. In Fig. 2, a list of the most often employed Vs is provided.

The elements that served as the foundation for this article's documentary are discussed in this part. For our survey to be conducted in a methodical manner and to identify the survey's subject, inclusion and exclusion criteria were developed to indicate which features would be valued and which would not. Using relevant keywords like "Big Data," "Machine learning," "Data mining," "disease diagnosis," "image

processing,” “Covid disease,” “deep learning,” “prediction,” and “detection,” among others, a thorough search was conducted using popular scientific bibliographical databases as PubMed, Clarivate Web of Science (WoS), and Google Scholar. Here, the inclusion and exclusion criteria were discussed henceforth we could acquire a substantial number of studies for analysis. To determine if the study was appropriate for inclusion, the abstracts of the publications that were searched were carefully examined. The studies related to big data analytics and its application to several domains were considered suitable for inclusion in this comprehensive study. The whole research articles were downloaded after choosing the relevant studies. This systematic literature analysis does not take into account secondary reports like short communications, non-peer-reviewed letter, editorials, and news articles. The selected works are published between 2015 and 2022. Research papers presented at renowned, highly referenced conferences in the field as well as in peer-reviewed publications were also included.

## 2.2 Class Imbalance Problem Associated with Big Data

At a rate never previously witnessed, databases are expanding in size and complexity, resulting in ever-larger datasets is defined as Big Data. The abundance samples for the various classes could not be balanced, which is a typical difficulty for categorization, especially with Big Data. Due to this prejudice in favour of the dominant class and contempt for the minority one, imbalanced categorization was first established a number of decades ago. Despite the fact that there are more imbalanced classification algorithms than ever before, they are still primarily concerned with small datasets rather than the new reality of big data. Johnson et al., [34] have addressed the class imbalance problem in training predictive models. To assess data sampling techniques for addressing high class imbalance the authors examined three data sets of varied complexity using deep neural networks based big data analytics. Although this issue has been extensively researched, the big data era is bringing up more extreme degrees of imbalance that are getting harder to forecast. In the year 2021, Juez et al. [35] have carried out the experimental analysis using ensemble classifier and resampling models which is conducted on imbalanced Big Data. The entire implementation is performed on spark clusters and compares with the Bayesian model in terms of classifier performance and time consumed. By interlinking Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) for the attack detection system, Al et al. [36] created the Hybrid Deep Learning (HDL) technique. To lessen the impact of data imbalance, Synthetic Minority Over-sampling Technique (SMOTE) sampling techniques along with Tomek-Links was employed. The Google Colab environment used PySpark, which gives Python provision on the Apache Spark platform. The CIDDs-001 data set served as the basis for the model’s multiclass assessment, while the UNS-NB15 data set served as the basis for the model’s binary classification assessment. An automated method for the detection of financial fraud, that connected to credit card transactions, has been demonstrated by Gupta et al. [38]. In order to predict fraud, the authors introduced

machine learning models such as random forests, support vector machines, logistic regression, and naive bayes on big data.

An assorted bagging-based stacked ensemble model was demonstrated by Sobanadevi et al. [40] in context of fraud detection while proceeding credit card transaction. The approach is made up of a bagging strategy that tries to deal with heterogeneous base learners and data imbalance problem that serve as decision rule extraction methods on complicated data. Even while banks always work to increase security to avoid fraud, sometimes criminals are able to get beyond these measures. Johnson et al. [41], have developed Deep Neural Network based on Bayesian a posteriori probability for defining threshold strategies on training the imbalanced dataset. The best parameter method uses training or validation of information to discover the threshold of classification that exploits the geometric mean. Prior probability of the positive class has been used as a classification threshold in the Prior threshold technique, which eliminates the need for optimization. To take into consideration random error, several designs are investigated, and all tests are performed 30 times. The optimal threshold and the positive class prior exhibit a substantial correlation, according to linear models and visualizations. Adaptive synthesis reliant on big data analytics has been developed by Javaid et al. [42] for the detection of electricity theft (ETD). The authors developed a deep Siamese network by combining LSTM and CNN to distinguish fraud and genuine customers. In particular, the duty of parameters mining from weekly energy usage patterns is assigned to a CNN component, whereas the LSTM component handles the sequential knowledge. The DSN then applies the final decision after considering the shared characteristics offered by the CNN-LSTM.

Data source from the State Grid Corporation of China (SGCC), which contains information on power use over 1035 days for two classes: regular and fraudulent, was used by Arif et al. [43]. This study proposes a methodology for detecting electrical theft that involves the following four steps: smoothing, data balance, extraction of features, and categorisation. The missing values are replaced with the help of smoothing process, then resampling approach was executed. To classify normal and theft data, residual network was employed to extract related features and are subjected to the corresponding learning algorithm presented. Arif et al. [44] created the Tomek Link Borderline Synthetic Minority Oversampling Technique with Support Vector Machine (TBSSVM) and Temporal CNN with improved Multi-Layer Perceptron for the goal of detecting energy theft (TCN-EMLP). The former evenly distributes the test dataset's occurrences of the majority and minority classes. The former distinguishes between genuine and deceptive consumers. Furthermore, the assignment of various weights causes deep learning models to have large volatility in their final outcomes. In order to lower the large variance, an average ensemble method was used on Pakistan Residential Electricity Consumption (PRECON) and State Grid Cooperation of China (SGCC) datasets. Hou et al. [45] have developed unified model to lessen the data imbalance issue. The author suggested a two-stage training method using self-attention-based time-varying forecast approach. In order to mine frequent patterns from time series, the authors utilized an encoder-decoder component with the multi self-attention technique. Then, in order to improve the outcomes of specific periods and address the imbalance problem, the author suggested

a time-varying optimization approach. Furthermore, to emphasize the significance of similar historical values to foresee outcomes, the authors suggest inverse spatial interest in place of conventional attention.

### 2.3 Big Data Characteristics in Sentimental Analysis

Numerous types of programs employ big data to extract important information that is used to make business choices, monitor certain behaviours, or identify potential security threats. Sentiment analysis (SA) is the most active research disciplines dependent on big data, despite the fact that their involvement would be very advantageous. To perform sentimental analysis on the gathered stockTwits dataset, Sohangir et al. [28] developed the deep learning technique. In the classification of opinion mining on stock market data, doc2vec was utilized together with the incorporation of LSTM and CNN. Hassib et al. [29] have focused on data imbalance problem in handling financial big data. Optimization approach was developed using data mining techniques, whereas local optima problem was resolved. Here, pre-processing step was carried out with the help of LSH-SMOTE algorithm. Then, grey wolf optimization was adopted to the bidirectional recurrent neural network in enhancing the global optimum solution. In the network big data environment, Liu et al.'s [30] comment texts for emotional analysis were initiated in the year 2020. The model being demonstrated was created using the combined Bag of Words (CBOW) language and the deep learning methodology. The feedforward neural networks were used to produce the vector representation of text. The Convolutional Neural Network (CNN), which was trained using the labelled training set, then captures the text semantic characteristics. The Softmax classifier of traditional CNN was the final component to add the Dropout approach, which can successfully avoid the model from over-fitting and has enhanced classification performance. Zhai et al. [31] developed Multi-Attention Fusion Modeling (Multi-AFM), which combines global and local attention to produce a believable contextual representation via gating unit control. Hybrid Lexicon-Naive Bayesian Classifier (HL-NBC) technique for sentiment analysis was explored by Rodrigues et al. [32]. Additionally, subject categorization, which divides tweets into several groups and blocks out unnecessary comments, occurred before the sentiment analysis algorithm. Lau et al. [33] created a parallel aspect-oriented sentiment analysis technique to glean customer insights from a substantial number of online product reviews. The suggested architecture puts into practice a broad empirical assessment of a co-evolving extreme learning machine-supported sentiment-enhanced sales forecasting method.

### 2.4 Big Data Analytics in Handling Financial Crisis

Budgetary credit rating is one of the most severe threats financial institutions must face financial institutions must deal with; however, with these different data sets, typical predictive algorithms and classifiers could not be fair or reliable enough to assess credit risk. Due to the enormously rising financial information from various sources like big data. Big data mining has been studied by Zhou et al. [21] in the



financial and banking industries for risk management prevention. Particle Swarm Optimization (PSO) was used to enhance a Backpropagation (BP) neural network in the model that is being presented in order to categorize financial risks. On the dataset of on-balance sheet and off-balance sheet items, Hadoop HDFS and Apache Spark are combined and the entire system operates in a nonlinear parallel optimization approach. A data-driven operation framework was introduced by Zhang et al. [22] which combines the framework with a time series data streaming massive data infrastructure and extricates the relationships of complicated significant components. The QuantCloud platform is the name of this integrated system was implemented on New York Stock Exchange (NYSE) data. In essence, QuantCloud manages massive volumes of diverse market information in a data-parallel manner while carrying out extensive event processing in a data-driven manner. Data cleaning process was taken place and the data modelling was carried out using autoregressive-moving average (ARMA) method. The research framework for a system providing regional financial data services in a smart financial context was designed by Wensheng et al. [23]. From three perspectives—functional scope, service model, and operating principle—it explores the underlying condition of building a rural financial information service platform in China and draws advancement conclusions from the current state of the platform's development and deployment there. To apply the graph embedding technique, an intelligent and distributed Big Data system was developed by Zhou et al. [24] for identifying financial fraud on the Internet. Node2Vec was created to acquire and express the spatial information in the commercial graph structure into low-dimensional sparse matrices, enabling the effective identification and prediction of data samples of the large-scale dataset with the deep neural network.

To handle the big dataset in simultaneously, the method was spread and carried out on clusters of Apache Spark GraphX and Hadoop. For Big Data analytics, Dos et al. [25] presented a data distribution strategy made up of hybridized Cloud Computing and Volunteer Computing infrastructures. A model for resource allocation in hybrid infrastructure and services was developed, as well as an HR Alloc Method for choosing where to deposit data in big data applications. In order to study three underlying patterns, namely investing pedigrees, investment firms, and structural gaps, Yang et al. [26] presented rapid networking methodologies from financial big data. They first described a pedigree classification system to recognize financial pedigrees, which draws inspiration from disjoint collections and route compression. Second, by offering a pruning approach and a data format known as “2-tuple list,” researchers established a linear-time structure mining algorithm in the network (SMAN) for analysing investment businesses and hierarchical gaps from the investment pedigree. Ruan et al. developed three fuzzy correlation measurement techniques in [27]: the centroid-based portion, the integral-based measure, and the cut-based ration. Utilizing data from the Stock Price Index SSI and exchange rates of foreign currencies over the Chinese Yuan from 22 January 2013 to 17 May 2018, we evaluated the performance of our algorithms. More significantly, the SSI and these significant exchange rates were found to be directly correlated. The use of maximum, minimum, or terminal values in daily exchange rates and stock prices affects the significant Granger causality of exchange rates over SSI, but there was no opposite causality from SSI to exchange rates. While the Euro over the Chinese Yuan was



adversely associated with the SSI and was documented as a Granger source to the SSI with an implication level of 1 percent, the Hong Kong Dollar over the Chinese Yuan and the U.S. Dollar over the Chinese Yuan are positively connected with the SSI. The comparison of a wide range of applications related to big data paradigm is illustrated in Table 1.

## 2.5 Data Management Using Spark to Manage Big Data Streaming

Relying on the Apache Spark open-standard Big Data processing platform that was installed in the cloud and focuses on smearing machine learning models to streaming Big Data, a real-time remote health status prediction system was created by Nair et al. [68]. When someone tweets about their health in this system, the program instantaneously receives, extracts, and analyzes those features to estimate the user's health state. The user is then immediately contacted to take the necessary action. Kılınc et al. [70] have offered an automated approach for identifying fake news detection scheme. The machine learning library of Apache Spark employed the Nave Bayes technique for training and testing both sentiment classification and fraudulent account detection systems. Ramírez et al. [71] have developed incremental model based on nearest neighbour algorithm in case of applying to demanding schemes. The presented distributed classifier was executed in the Apache spark platform to stream the big data. The authors also provided a method of progressive instance selection for large data sources that modify and delete out-of-date instances from the particular instance continually. This lessens the original classifier's high processing demands, making it appropriate for the situation under consideration. Park et al. [72] have designed big data analytics in the spark framework where streaming was performed in memory-based cluster-based computing standard. Spark was offered by the open-source ASF (Apache Software Foundation) community.

Carcillo et al. [73] developed a Scalable Real-time Fraud Finder (SCARFF) using a substantial dataset of real credit card transactions. SCARFF integrates Big Data technologies (Kafka, Spark, and Cassandra) with a machine learning technique that handles asymmetry and reinforcement delay. Rathore et al. [74] proposed a real-time Big Data stream computing method by implementing the Hadoop MapReduce comparable models on graphics processing units (GPUs). Spark with GPU was paired with a parallel and distributed Hadoop ecosystem environment to increase the system's power and ability to handle the enormous volume of high-speed streaming. By breaking large Big Data records into fixed-size blocks, a MapReduce-like method for GPUs was created to calculate statistical parameters. Spark Streaming-based system for monitoring online Internet traffic was suggested by Zhou et al. in [75]. The collector, message system, and stream processor are the three components that make up the system. The TCP performance monitoring was chosen by the authors as a unique use case to demonstrate how their suggested method might be used for network monitoring. A distributed technique called SWEclat was suggested by Xiao et al. [76] for mining frequent item sets across enormous streaming data. The program stores the dataset in a vertical data structure and processes streaming data using a sliding window. In order to split these RDDs for distributed processing, this

**Table 1** Comparison of a wide range of big data applications

Author and year	Category	Technique	Limitation
Johnson et al. [34], 2020	Issue with Class Imbalance in Big Data Analytics	Deep neural network	Due to class imbalance problem, it is harder to forecast
Juez et al. [35], 2021		Ensemble classifier (Bagging and Boosting)	Dissimilar resampling ratio, affects ensemble classifier
Al et al. [36], 2021		Hybrid Deep Learning (HDL)	Big data's class imbalance nature lowers the effectiveness of the current classification techniques
Gupta et al. [38], 2021		Machine learning models include Naive Bayes, Random Forest, Logistic Regression and SVM	Handcrafted features frequently perform less accurately in terms of categorisation
Sobanadevi et al. [40], 2021		Heterogeneous bagging-based stacked ensemble model	On the basis of a massive data class imbalance dataset, the calculated False predictive rate was insufficient for classifying fraud detection
Johnson et al. [41], 2021		Deep Neural Network based on Bayesian a posteriori probability	Across a variety of data sets, there is no link between the minority class size and the ideal choice threshold that performs as well as the model that is provided
Javaid et al. [42], 2021		LSTM and CNN and integrated deep Siamese network (DSN)	The presented deep learning methodology does not report the problem of class imbalance
Arif et al. [43], 2021		Decision trees (DT), random forests (RF), and adaptive boosting are examples of tree-based classifiers (AdaBoost)	During data sampling, time consumption is very high for data preparation in the presented model
Arif et al. [44], 2022		Tomek Link Temporal Convolutional Network Enhanced Multi-Layer Perceptron with Support Vector Machine Borderline Synthetic Minority Oversampling Technique	Since the suggested technique falls short of obtaining more specific information, the detection rate was somewhat impacted
Hou et al. [45], 2021		Encoder-decoder component with the multi self-attention technique	Need improvement in accuracy measure for long-term forecasting

**Table 1** (continued)

Author and year	Category	Technique	Limitation
Sohangir et al. [28], 2018	Sentimental analysis under big data analytics	Convolutional neural networks, doc2vec, and long short-term memory	The measure of accuracy was comparatively less
Hassib et al. [29], 2019		Bidirectional Recurrent Neural Networks, Grey Wolf Optimizer	The massive data set's running time was a little bit excessive due to the numerous attributes
Liu et al. [30], 2020		Combining Convolutional Neural Network with Bag of Words	Despite the fact that the provided work was applied to data sets in both Chinese and English, the classification accuracy was not up to standard level
Zhai et al. [31], 2020		Multi-Attention Fusion Modelling (Multi-AFM)	The task-specific module is intimately connected with the pre-trained network, which makes the training process expensive and time-consuming
Rodrigues et al. [32], 2019		Hybrid Lexicon-Naive Bayesian Classifier (HL-NBC)	Cross-lingual sentiment categorization is a difficult problem since there isn't support for sentiment dictionaries
Lau et al. [33], 2018		Algorithm for parallel aspect-oriented sentiment analysis	A parallel model with a learning technique was required to improve the accuracy of aspect extraction from large amounts of data

**Table 1** (continued)

Author and year	Category	Technique	Limitation
Zhou et al. [21], 2019	Big Data Analytics for Financial Crisis Management	A neural network based on backpropagation and particle swarm optimization	In high-dimensional space, it is simple to fall into a local optimum and the iterative process has a low convergence rate
Zhang et al. [22], 2018		Autoregressive moving average driven by data (ARMA)	Estimating the ARMA parameters is still a difficult task
Wensheng et al. [23], 2020		Platform for providing financial information to rural areas in a smart financial environment	The system is insufficient for analysing information from multiple structures and high complexity because of the complicated resources and diversified structures
Zhou et al. [24] 2021		Clusters of Apache Spark GraphX and Hadoop financial fraud detection	Lacks in learning effectual set of features
Dos et al. [25], 2020		HR_Alloc Algorithm with hybrid infrastructure	Transfer speed is high rather than traditional approaches
Yang et al. [26], 2020		Linear-time structure mining algorithm in network (SMAN)	Parallel graph clustering approaches was suggested in order to maintain big data
Ruan et al. [27], 2019		Big data fuzzy correlation measuring techniques with applications to stock prices and currency rates	These systems' accuracy is impaired since algorithms use incorrect data and inputs

method was created by Apache Spark and stores streaming data and datasets in vertical data format using Spark RDD.

## 2.6 Map Reduce-Based Method in Different Streams of Big Data Environment

A weighted distributed long short-term memory archetypal called WND-LSTM was published by Xia et al. [46] relying on a MapReduce parallel computing environment. More specifically, the Hadoop distributed computing platform developed a decentralized forecasting method for anticipating traffic flow using MapReduce to address the challenges with storage and computation associated with managing massive amounts of traffic flow data with an independent learning approach. Bawankule et al. [47] introduced the Historical data based Reduce tasks scheduling (HDRTS) method to decrease the operational skew caused by the Reduce phase of the MapReduce process. To dynamically manage intrusion detection, Asif et al. [48] demonstrated the MapReduce-Based Intelligence Mechanism for Intrusion Detection (MR-IMID). The suggested MR-IMID reliably processes large data sets using technology that is readily available. In-depth research on task scheduling in diverse environments was provided by Wang et al. [49], who also suggests the task scheduling algorithm HTD. Pandey et al. [50] have developed a deadline-aware, energy-efficient MR scheduling problem for the Hadoop YARN architecture. The time-indexed binary decision variables to describe the scheduling issue under consideration was analysed as an integer program. Then, using the knowledge that tasks have varied energy usage values on various computers, to schedule map and decrease tasks on the heterogeneous cluster machines, a heuristic technique is developed.

The study by Baruah et al. [51] focused on the deep neuro-fuzzy network technique for forecasting student performance that is based on fractional competitive multi-verse optimization. The mapper and reducer stages of the MapReduce framework were developed in order to execute the performance of student prediction approach utilizing the deep learning classifier. Narayana et al. [54] have developed a classification solution using the Ant Cat Swarm Optimization-enabled Deep Recurrent Neural Network (ACSO-enabled Deep RNN) via Map Reduce framework, which integrates the Ant Lion Optimization strategy with the Cat Swarm Optimization technique. Massive data categorization and feature selection are carried out using the Map Reduce framework. The features are chosen via fuzzy clustering based on black hole entropy and Pearson correlation. In the reduction part, categorization is done with the aid of a Deep RNN that was trained using a custom ACSO algorithm. The Gaussian Relevance Vector MapReduce-based Annealed Glowworm Optimization Scheduling (GRVM-AGS) hybrid model, developed by Patan et al. [52], was intended to progress the distribution of big medical data files across several clinicians while needing less time and more efficient scheduling. A GRVM approach was initially established for the predictive analysis of incoming medical data.

Arunadevi et al. [58] have described a Cuckoo Search Augmented Map Reduce for Predictive Scheduling (CSA-MRPS) system. Foresight research applied MapReduce processes on discretized data using Multi-Objective Ranked Cuckoo Search

Optimization (MRCSA). Every sort of area's pure spectra and perturbed/mixed spectra are qualitatively recorded using the extended extreme learning machine (IELM) technique, which was created by Roy et al. [56]. The Jahazpur mineralized belt has been extensively mapped using a MapReduce model that combines the IELM technique and AVIRIS-NG (Airborne Visible-Infrared Imaging Spectrometer-Next Generation) observation. Mapreduce (MR) model has been used to the huge data of maritime navigation by Pham et al. [57]. Specifically, the author uses a popular clustering method called K-means, which is based on the MR model, to analyze the data of marine traffic in the South Vietnam Sea region. Ramsingh et al. [55] propose the NBC-TFIDF (Naive Bayes Classifier—Term Frequency Inverse Document Frequency) technique and a MapReduce-Based Hybrid. Using a Map Reduce-Based Hybrid NBC, the data are categorized according to the polarity score assigned to each sentence in the social media data. Chawla et al. [53] have presented MuSe, an efficient distributed RDF storage solution for loading and querying RDF data using Hadoop MapReduce.

Using the ACO-GA method with HDFS map-reduce, Kumar et al.'s proposal for an enhanced query optimization procedure in big data (BD) is presented in [59]. The dataframe is initially pre-processed by utilizing the SHA-512 method to get the hash value (HV) and the HDFS map-reduce function to eliminate redundant data. The optimized query followed using ACO-GA algorithm. For the map-reduce framework-based training of deep belief networks (DBNs), Agarwal et al. [60] proposed two simultaneous models. For the layer-by-layer training of DBNs in both models, which followed the positive and negative phases, we employed several computers. Studies using data sets from the Toronto Emotional Speech Set and the Ryerson Audio-Visual Database of Emotional Speech and Song have revealed that the first proposed model, the First Parallel Map-Reduced-Based Deep Belief Network, is effective (FParMRBDBN). The hybrid Database-MapReduce system proposed by Pang et al. [61] would combine the benefits of both systems. The query optimizer AQUA+ that the authors proposed was made specifically for the hybrid system. Maheswari et al. [62] developed a Kernelized Spectral Clustering based Conditional Maximum Entropy MapReduce (KSC-CMEMR) method for reducing dimensionality.

## 2.7 Big Data in Healthcare Diagnosis

Big data analytics in healthcare follow a structure that is largely the same as traditional healthcare informatics. Independent systems with installed advanced analytics were frequently utilized in healthcare initiatives. Distributed processing is used when computation needs to be implemented across a number of locations owing to “big data.”

### 2.7.1 Big Data Application for Heart Disease Diagnosis

Thanga et al. [63] offer a huge health implementation model based on optimum artificial neural network (OANN) for the diagnosis of heart disease, which is regarded

as the worst disease in the entire globe. The two main parts of the proposed OANN are the distance-based misclassified instance removal (DBMIR) approach and the teaching and learning-based optimization (TLBO) method for ANN (TLBO-ANN). The conceptual system is developed using a Big Data framework such as Apache Spark. The offline forecasting stage and the online forecast stage are how the presented OANN model functions. Ed et al. [64] established a real-time cardiac disease predicting framework that is based on Apache Spark, a potent large-scale distributed computing platform that can be successfully used to process streaming data events as opposed to machine learning employing in-memory operations. The technology is made up of two basic components: data storage and visualisation and streaming processing. The first predicts cardiac illness by applying a classification model to data events using Spark MLlib and Spark streaming. The significant quantity of created data is stored by the seconds using Apache Cassandra.

According to Vaishali et al. [65], a significant collection of medical records is employed as input. It is meant to extract the required data from the records of heart disease patients in this particular dataset using the map reduction approach. Heart disease [115, 116] is a severe health problem and the leading cause of death worldwide. Early heart illness diagnosis has become very important in the realm of medical research. The RR interval, QRS interval, and QT interval are a few features that are looked at to help diagnose heart illness. In the testing phase, the map reduction methodology is utilized to identify the disease and minimize the dataset after the classification algorithm determines if the patient is normal or abnormal. Rastogi et al. [66], to predict a patient's risk of getting heart disease, input factors such their gender, cholesterol, blood pressure, TTH, and stress can be taken into account. Data mining (DM) techniques such Naïve Bayes, decision trees, support vector machines, and logistic regression are studied using the heart disease database. Nayak et al. [67] emphasize the importance of early diagnosis of cardiac abnormalities in order to successfully treat these illnesses. This paper employs a number of data mining classification techniques, such as Decision Tree classification, Naïve Bayes classification, Support Vector Machine classification, and k-NN classification, to identify and take precautions against diseases at an early stage, so they can be treated and prevented.

### 2.7.2 Kidney Disease Classification with the Help of Big Data

Sisodia et al. [82] propose a healthcare multi-phase architecture (HCMP) for predicting chronic renal illness. Data collection, data storage, management, processing, analysis, and report preparation are the six layers of the HCMP architecture. The data-storage and data-management layers were implemented on a heterogeneous Hadoop cluster, and the profiling approaches were used to account for three situations in order to ascertain the capacity ratio of each DataNode in the cluster. The MySymptom algorithm and MapReduce, which is used for parallel data processing, have been used to filter the renal dataset of patients according to their symptoms at the data processing layer. Diez et al. [83] focuses on an area of medicine called kidney transplant therapy for end-stage renal illness, where machine learning is starting to be utilised as a supplemental resource to estimate or make judgments. Abdelaziz



and others [84], utilised CKD prediction accuracy in a cloud setting. The stakeholders in health care in smart cities believe the cloud-IOT-based prediction of hazardous illnesses like CKD to be a significant challenge. Chronic kidney disease (CKD) prediction is given as an example of a health service offered using cloud computing. Cloud computing enables residents of smart cities to estimate the presence of CKD at anytime and anywhere. By combining two intelligent techniques, namely linear regression (LR) and neural network, this research provides a hybrid intelligent model for predicting CKD based on cloud-IOT.

### 2.7.3 Brain Tumour in Big Data Applications

Mansour et al. [85] introduce an artificial intelligence and big data analytics-based ICH e-diagnosis (AIBDA-ICH) model using CT images. IoMT devices are used in the given model's data collection process. In the presented AIBDA-ICH technique, the damaged areas in the CT scans are located using a graph cut-based segmentation model. The main purpose of the Hadoop Ecosystem and its parts is to manage massive amounts of data. The capsule network (CapsNet) model is also employed as a feature extractor to obtain a useful collection of feature vectors. Last but not least, a fuzzy deep neural network (FDNN) model was used for classification by the aforementioned AIBDA-ICH approach. The improved results of the AIBDA-ICH technique were validated through a large number of simulations, and the outcomes are examined from a variety of perspectives. Big data and image processing have been combined for the segmentation and classification of brain tumours by HS et al. [86]. Using the MATLAB Hadoop system, a big data analysis is performed on the image of the brain tumor. The BraTS dataset is delivered to the Hadoop and Matlab Distributed Computing Server (MDCS) system for processing, which is performed by one master node and four slave nodes (multimode) in an MDCS configuration. The findings of this inquiry are broken down into its component pieces using the dual-tree complex Gabor wavelet transform (DTCGWT). Malignant and benign brain tumors are distinguished from each other using the generated feature vectors by the CNN model. If a malignant brain tumour was found, the image will be segmented using the fuzzy level set method based on the manta ray foraging algorithm (FLSM-MRF).

### 2.7.4 Diabetic Disease Identification with the Help of Big Data Analytics

To assist the healthcare system in making speedy and informed judgments, Saluja et al. [69] used MapReduce. Every day, a tremendous amount of data is generated in the life sciences. Using MapReduce, the vast amount of clinical data is divided into a variety of categories, making it easy to understand for future use. The major goal is to filter different aspects of diabetes and heart disease while using clinical data to build healthcare information for designing a health service that is physician. Subramaniyan et al. [77] used big data with Large-scale computers and machine learning to provide predictive analytics for extracting inherent information. A prerequisite for Big Data computing, cloud computing has arisen as a service-oriented computing architecture for processing enormous amounts of

quickly expanding data at a quicker scale. Big data frameworks like Hadoop and Spark can be utilized in conjunction with machine learning approaches. AlZubi et al. [78] use big data and classification techniques, such as effective map reduction technologies, to find diabetes. The map reduction concept was utilized to effectively generate the little piece of data after the data were initially composed from a large dataset. Following this, the acquired dataset is normalised to remove any noise that was present. The ant bee colony strategy, which makes advantage of ant traits like wandering, was then used to choose the statistical attributes. Support vector machine with multilayer neural network was employed to train the chosen features.

In the Hadoop framework, Hatua et al. [79] have developed a Diabetic Retinopathy (DR) detection approach that is quick and accurate and can detect the first indications of diabetes from retinal images of the eye. According to the suggested method, there are five levels of diabetic retinopathy: Proliferative DR, Mild DR, Moderate DR, Severe DR, and No DR. For the purpose of categorizing images of diabetic retinopathy, the suggested method separates feature extraction, feature reduction, and image classification into three distinct procedures. Each of the images for diabetic retinopathy is represented by the Histogram of Oriented Gradients (HOG) at the beginning of the method. Principal Component Analysis is employed to diminish the dimension of HOG characteristics. The method's final step is a classifier called K-Nearest Neighbours (KNN). Sivakumar et al. [80] presented the Equidistant Heuristic and Duplex Deep Neural Network (EH-DDNN) technique for the early detection and prognosis of diabetic illness. First, the feature selection method known as Equidistant Heuristic Pruning (EHP) is discussed using the Big Data dataset as input. The EHP separately divides the incoming data matrix into rows and columns. EHP divides neighbourhood assessments while dropping communication overhead and time, considerably boosting computation correlation. It does this by utilizing conditional non-alignment assessment and heuristics methodologies. The result is fewer features that are better for early prediction and the elimination of redundant and unneeded attributes. A Duplex Deep Neural Network (DDNN) is then developed for early assessments by fusing the intrinsic characteristics with nonlinear processing features and linear response. The efficiency of classification and clustering methods was evaluated for diabetic medical data by Mamatha and colleagues [81] which is detailed in Table 2.

## 2.8 Data Mining Approaches in Dealing with COVID Crisis

Big data presents a huge prospect for physicians, epidemiologists, and experts in health policy to make selections based on the best available knowledge, thereby improving patient care. Big data is not just a contemporary reality for researchers and scientists; it is also a requirement that must be correctly comprehended and used in the search for new knowledge. Big data will be essential to managing the COVID-19 epidemic in today's digital environment. The generic progression of COVID detection and prediction techniques is offered in the following figure.

**Table 2** A healthcare application from the perspective of big data analytics

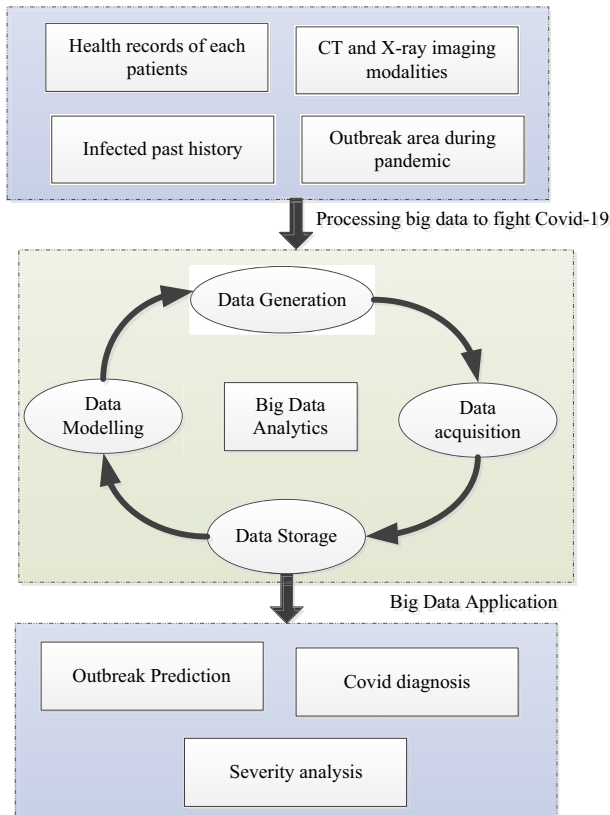
Author and year	Disease category	Findings	Dataset	Limitation
Thanga et al. [63], 2021	Heart disease diagnosis under big data analytics	The ideal big data application for heart disease diagnostics uses artificial neural networks	The UCI repository's benchmark heart disease dataset	Computation time is high. Also, the performance measure is not well obtained due to the lack of feature selection model
Ed et al. [64], 2019		Using big data and artificial intelligence to detect heart problems early	The UCI repository's dataset	Execution time was very much higher
Vaishali et al. [65], 2016		Map Reduce approach is used in big data analysis for heart disease diagnosis systems	The UCI repository's dataset	The present disease is only tested with binary classification model
Rastogi et al. [66], 2020		A big data and IoT application and analysis using machine learning for intelligent heart disease prediction	Cleveland dataset	Lack in selecting optimal set of parameters
Nayak et al. [67], 2019		Using a Big Data analytical tool, a comparative analysis of algorithms for classifying heart disease	Compared different dataset	Not covered an extensive number of articles

**Table 2** (continued)

Author and year	Disease category	Findings	Dataset	Limitation
Sisodia et al. [82], 2022	Kidney disease using big data analytics	Using Kernel distributed Nave Bayes and Healthcare Multi-phase Architecture (HCMP), one can forecast the development of chronic kidney disease	Sample of scored chronic kidney dataset	Need improvement in accuracy measure for the diagnosis of chronic kidney disease
Diez et al. [83], 2021		Examining the effects of artificial intelligence and big data on the management of end-stage renal disease		Overfitting is the major issue occurring in the proposed ensemble machine learning approaches
Abdelaziz et al. [84], 2019		Hybrid intelligent model with linear regression (LR) and neural network (NN) with big data		The obtained error function is not minimum even though using hybrid approach
Mansour et al. [85], 2021	Brain tumour classification by means of big data analytics	The capsule network (CapsNet) model is employed as a feature extractor, while the fuzzy deep neural network (FDNN) model is utilized for classification	CT images	Hyperparameter tuning strategies were missing in the presented design which lacks system performance
HS et al. [86], 2022		Brain tumour recognition method based on dual-tree complex Gabor wavelet transform and neural networks using Hadoop big data analysis	BraTSDataset	Processing time is slow due to the integration of Hadoop framework

**Table 2** (continued)

Author and year	Disease category	Findings	Dataset	Limitation
Saluja et al. [69], 2021	Diabetic detection with big data applications	Big Data Analysis of Diabetes and Heart Disease Using the MapReduce Framework	The 45–54 age group has diabetes, and comes with heart-related conditions	The presented model deals with the unstructured data which degrades the system performance
Subramaniyan et al. [77], 2020		Employing big data analytics, a semi-supervised machine learning algorithm may predict diabetes	Diabetic big dataset	The automatic generation of features are rare while using machine learning algorithm which in turns omit appropriate features
AlZubi et al. [78], 2020		Diabetes patients employing map reduction and classification methods in big data analysis		The categorization outcome is not successfully attained
Haua et al. [79], 2021		Big data in the Hadoop architecture for the early diagnosis of diabetic retinopathy	Benchmark dataset	Time consumption is high as comparative to another existing model
Sivakumar et al. [80], 2021		In order to predict the progression of diabetes, an IoT-based big data platform employs equidistant heuristic and a duplex deep neural network	Diabetes Data Set from the UCI repository and the Pima Indians with Diabetes Disease dataset	Due to sophisticated data models, training is quite costly. Deep learning also needs hundreds of workstations and pricey GPUs
Mamatha et al. [81], 2019		The analysis and detection of diabetes is one application of big data in healthcare	The medical dataset on diabetes is sourced from the UCI machine learning repository	It is suggested to perform ensemble classification in future henceforth the performance of the system will also be improved



**Fig. 3** General architecture of COVID diagnosis using big data analytics

Figure 3 depicts the fundamental architecture of COVID diagnosis using big data analytics. In this section, we conduct a literature review to highlight the significance of various research in the area of COVID-19-based big data analysis.

### 2.8.1 COVID Diagnosis Using Machine Learning

Big data applications are essential for managing pandemic circumstances, such as predicting COVID-19 outbreaks and locating COVID-19 cases and patterns of transmission. The textual clinical report was categorized using a machine learning (ML) method according to Ramanathan et al.'s [94] documents mining approach from 1994. The ensemble ML classifier technique was used to abstract features from the term frequency-inverse document frequency (TF/IDF) advanced approaches, a productive way of information retrieval from the Corona dataset. The three-way COVID-19 steps are divided into categories based on how the feature was extracted using a machine learning-based information retrieval approach. The TF/IDF for coronavirus classification and prediction is developed to quantify and statistically

analyze the text data mining of the COVID-19 patient record list. The deep learning neural network-based method nCOVnet developed by Panwar et al. [96] may be used to swiftly screen for COVID-19 by evaluating the X-rays of patients and looking for visual indicators seen in the chest radiography imaging of COVID-19 patients.

Twitter data was analysed using the R programming language, according to Kaur et al. (1997). We have gathered Twitter information based on hashtag terms like COVID-19, coronavirus, fatalities, new case, and recovered. The Hybrid Heterogeneous Support Vector Machine (H-SVM) algorithm was developed in this study to categorize sentiment ratings into three groups: neutral, negative, and positive. Wang et al. [98] examine the feature selection algorithm of big data samples in order to select typical economic indicators from among the numerous economic statistics of the sports industry and show the growth trend of the sports industry. A deep learning approach based on large data feature selection is advised. Deep learning and data fusion are used after initially developing a framework for huge data feature selection. This strategy incorporates a number of forward-looking elements as well as a particular reference point for data on the evolution of the sports industry.

## 2.8.2 COVID Diagnosis Using Deep Learning

Big data technologies are a vital tool in the war against COVID-19 in a range of enticing applications, including as pandemic monitoring, viral sensing and therapy, and diagnostic support. In order to simulate the increase rate in the number of covariates, Chew et al. [87] state that a hybrid deep-learning model known as ODANN has been built to handle daily COVID-19 time-series records and massive quantities of COVID-19 related Twitter data at the same time. The neural networks (NN) that make up this model are combined with an analytical framework and feature extraction techniques. Elghamrawy et al. from [88] offers a comprehensive analysis of the part Deep Learning and Big Data Analytics play in disease prevention. Additionally, a model (DLBD-COV) based on Big Data analytics and influenced by H2O's Deep Learning is recommended for early patient detection using X-ray and CT images. The scalable handling machine learning framework is used to build the proposed diagnostic model (H2O). CNNs and generative adversarial networks (GAN) perform classification in contrast to each other. The remedy proposed by Jamshidi et al. [89] uses artificial intelligence to combat the pathogen (AI) [117].

Several Deep Learning (DL) methods have been shown to accomplish this goal, including the Extreme Learning Machine (ELM), Long/Short-Term Memory, and Generative Adversarial Networks (GANs). It describes an integrated bioinformatics approach where different informational elements from a variety of structured and unstructured data sources are combined to develop platforms that are practical for academics and medical professionals to use. Oh et al. [90] recommend employing a patch-based convolutional neural network technique with a manageable amount of trainable parameters for the diagnosis of COVID-19. The methodology was based on our statistical analysis of the possible imaging biomarkers identified in the CXR radiographs. Using deep learning and laboratory data, Alakus et al.'s clinical prediction models [91] identify the patients who are most likely to experience a COVID-19 illness. The precision, F1-score, recall, AUC, and accuracy scores of our models



were used to evaluate their predictive ability. According to Awan et al.'s [99] considerable data framework, a Deep Transfer Learning (DTL) technique using Convolutional Neural Network (CNN) three architectures—InceptionV3, ResNet50, and VGG19—was deployed using COVID-19 chest X-ray images. The three models are correctly evaluated in the COVID-19 and conventional X-ray image classes [118].

### 2.8.3 Influence of COVID-19 Epidemic Situation Based on Big Data

Luo et al. [92] exhibited deep learning methods for assessing diners' evaluations of restaurant characteristics and locating reviews with discrepant ratings. The method of Prasanth et al. [93] for foreseeing future occurrences of infection is offered. This strategy is based on the analysis of data obtained from the search phrases of people in the affected region. The study practices data on COVID-19 spread from the European Centre for Disease Prevention and Control as well as pertinent Google Patterns of specific search phrases connected to the pandemic in order to estimate upcoming trends of daily new cases, cumulative cases, and deaths for India, the USA, and the UK. In order to do this, the network parameters of the Long Short-Term Memory (LSTM) network are enhanced using a hybrid Grey Wolf Optimizer (GWO)-LSTM model (GWO). Ghosh et al. [95] investigated data from Twitter microblogs to ascertain how COVID-19 impacts people's mental health, specifically depression (tweets). Recurrent neural networks are used in our special pipeline to accurately identify depressed tweets with a 99.42% accuracy rate (in the form of long-short term memories, or LSTM). The detailed view and impact of the COVID-19 epidemic in different streams collected since 2021 is illustrated in Table 3.

### 2.8.4 Impact of Imaging Modalities in COVID Diagnosis Using Big Data Analytics

Wang et al. [100] developed COVID-Net, a deep convolutional neural network architecture that is available on the internet source, to recognize COVID-19 instances from chest X-ray (CXR). The most COVID-19-positive instances are found in COVIDx, an open access benchmark dataset we developed, which consists of 13,975 CXR images from 13,870 patient cases. Yang et al. [101] used four powerful pre-trained CNN architectures: VGG16, DenseNet121, ResNet50, and ResNet152 for the COVID-19 CT-scan binary classification task. The Fast.AI ResNet framework will be used to establish the networks' ideal design, pre-processing, and training settings almost immediately. We also employed transfer learning techniques to get over the shortage of data and cut the training time. The binary and multi-class classification of X-ray imaging tasks were carried out using modified VGG16 deep transfer learning architecture. Ohata et al. [102] have created an automated COVID-19 infection detection system based on chest X-ray. The datasets produced for this experiment consist of 194 X-ray images of patients with coronavirus and 194 X-ray images of healthy people. The authors adopted the concept of transfer learning for this, because there are not enough photographs of patients with COVID-19 that are available to the general population. The scientists used several ImageNet-trained convolutional neural network (CNN) architectures and modified them to work as feature extractors for the X-ray images.

**Table 3** Impact of the COVID-19 epidemic in different streams collected since 2021

Author and year	Method used	Description	Dataset	Outcome	Limitation
Ramanathan et al. [94], 2021	Term frequency-inverse document frequency utilizing machine learning for textual big data mining	The ensemble ML classifier's algorithm extracted features from the Corona dataset using the effective information retrieval method TF/IDF	Original raw corona dataset stored at UCI	Accuracy 92.74%	The performance of manually created features is inferior to that of an automatic feature extraction method
Kaur et al. [97], 2021	Heterogeneous Hybrid Support Vector Machine (H-SVM)	A study of Twitter data based on hashtags. The sentiment categorization was carried out the authors using Hybrid Heterogeneous Support Vector Machine (H-SVM)	Twitter data about COVID-19 is available through an IEEE data port	Accuracy-96.3%	Sentimental analysis is carried out on a single language due to the lack of data availability
Chew et al. [87], 2021	Optimized data assimilated neural network	Using social media big data, along with hybrid deep learning, to foresee the spread of COVID-19	The massive collection of English-language Tweets, totaling millions, was taken from 244 open-source Source on Kaggle	MSE-0.00396	Data are collected only from social network analysis. It is suggested to perform COVID transmission from other sources also

Chowdhury et al. [103] proposed Parallel-Dilated COVIDNet, a parallel-dilated convolutional neural network (CNN)-based COVID-19 identification system from chest X-ray images (PDCOVIDNet). Before applying the suggested method to categorize the freely accessible chest X-ray collection, it must first be fully preloaded and improved. Convolution dilation rate variation is used as the proof-of-concept for using PDCOVIDNet to derive radiological features for COVID-19 identification. A precise framework for a convolutional neural network to differentiate COVID-19 patients from supplementary pneumonia cases was offered by Farooq et al. [107]. Awasthi et al. [108] have developed a compact, mobile-friendly, and successful deep learning algorithm for the detection of COVID-19 using lung US data. Three distinct courses were featured in this challenge, including COVID-19, pneumonia, and healthy. The developed network, known as Mini-COVIDNet, was compared to both lightweight and more sophisticated heavyweight neural network models. Hasan et al. [109] discuss the possibility of using convolutional neural networks to predict COVID-19 patients from a CT. The cutting-edge strategy for predicting COVID-19 is based on the most recent CNN architectural change (DenseNet-121). 2D Sparse Matrix Profile DenseNet has been found by Liu et al. [110] for the diagnosis of COVID-19 utilizing CT scans. According to Xiao et al. [111], a CNN with a parallel attention module (PAM-DenseNet) may perform well on coarse labels without manually specified infection zones. The dense connectivity structure achieves feature map reuse by adding direct connections from all previous levels to all following layers, and it may be able to extract representative features from fewer CT slices. The impact of imaging techniques in COVID diagnosis using big data analytics is described in Table 4.

### 3 Analysis from the State-of-the-Art

Data mining techniques were used to examine a huge number of publications for big data analytics, and 25 research that were connected to a certain set of COVID-related issues were chosen. The performance metrics has been derived on the basis of four basic metrics namely f-measure, specificity, accuracy and sensitivity. Among the total number of samples in the test, Accuracy represents the ratio of correct predictions. The evaluation undertaken by the previous studies for COVID diagnosis with and without big data analytics using data mining algorithms with the obtained performance measure is graphically displayed below.

The total number of classification model taken according to this survey paper is pictorially shown in the pie chart.

From the analysis, it describes algorithms is utilized utmost and used in current period for big data analytics. There are around 114 research articles are taken from learning models like DT, KNN, RF, NB, LR, ANN, DNN, CNN and SVM in the context of big data analytics. The vast majority of research methodologies used between 2015 and 2022 are based on deep learning techniques.

**Table 4** Impact of imaging techniques in COVID diagnosis using big data analytics

Author and year	Method used	Findings	Dataset	Outcome	Limitation
Panwar et al. [96], 2020	nCOVnet	X-ray scans of the person's lungs were examined to find the COVID-19	X-ray imaging dataset	Accuracy- 88%	Due to less input sample, overfitting issues are occurred in this presented system
Elghamrawy et al. [88], 2020	H2O's deep learning-inspired model	A big data analytics-based coronavirus disease model inspired by deep learning from H2O (COVID-19) diagnosis	Computed Tomography (CT) or X-ray imaging	Accuracy- 89.17	Due to the usage of real COVID dataset, training images are less which lessen classification accuracy
Jamshidi et al. [89], 2020	LSTM, generative adversarial networks (GANs), extreme learning machine (ELM)	Deep learning methods for COVID-19 diagnosis and treatment using AI	Set of COVID-related imaging modalities	–	Medical imaging and clinical data applications frequently have poor system performance
Oh et al. [90], 2020	Using patch-based convolutional neural networks	Utilizing Limited Training Data Sets, Deep Learning COVID-19 Features on CXR	Chest X-ray (CXR) image	Accuracy- 70.7%	Accuracy measure was not achieved up to the level
Alakus et al. [91], 2020	Deep learning-based COVID laboratory data	Deep learning was investigated to identify the patients who are most likely to acquire a COVID-19 condition	X-ray or CT images	Accuracy- 86.66%	The amount of the data is the main drawback of this study
Awan et al. [99], 2021	Using the deep transfer learning (DTL) method and three InceptionV3, ResNet50, and VGG19 convolutional neural network (CNN) architectures	COVID-19 in chest X-ray images detection: a deep learning method using big data	Chest X-ray images	Accuracy- 97%	To enhance system performance, TensorFlow integration into Apache Spark will be implemented in the future

**Table 4** (continued)

Author and year	Method used	Findings	Dataset	Outcome	Limitation
Wang et al. [98], 2020	Multilayer convolution neural network	The sports industry is evolving as a result of the COVID-19 pandemic situation using big data	Big data related to sport data in COVID situation	Accuracy-88.36%	It is discovered that the deep computing model's economic data categorization accuracy and flexibility have been significantly enhanced
Wang et al. [100], 2020	Tailored deep convolutional neural network	COVID-Net is a deep convolutional neural network created exclusively for identifying COVID-19 instances from chest X-ray images	Chest X-ray images	Accuracy 93.3%	As more information is gathered, next initiatives include enhancing sensitivity and PPV to COVID-19 infections
Yang et al. [101], 2021	VGG16, DenseNet121, ResNet50, and ResNet152 are pre-trained CNN models	Deep learning algorithms for COVID-19 detection and analysis in medical images	CT and X-ray scan images	Accuracy- 96%	The presented model lacks in performing pre-processing step prior to classification task
Ohata et al. [102], 2020	CNN Transfer learning	Transfer learning-based automatic COVID-19 infection identification	Chest X-ray images	Accuracy- 95.6%	The presented approach has not undergone on big data analytics for handling high volume dataset
Chowdhury et al. [103] 2020	A parallel-dilated convolutional neural network is PDCOVIDNet	An architecture for a parallel-dilated convolutional neural network that can recognize COVID-19 is called PDCOVIDNet	Chest X-ray images	Accuracy- 96.58%	Training images are less hence accuracy becomes less

**Table 4** (continued)

Author and year	Method used	Findings	Dataset	Outcome	Limitation
Canayaz et al. [104] 2021	Deep learning models like ResNet, GoogleNet, VGG19, and AlexNet. Grey wolf optimization in binary	Deep neural networks and feature selection based on meta-heuristics are used in MH-COVIDNet to diagnose COVID-19	X-ray images	Accuracy-95.71%	It is recommended to use big data analytics techniques to handle and manage dataset
Aboutalebti et al. [105], 2021	COVID-Net Convolutional 3D neural network	3D architectures of convolutional neural networks for evaluating the severity of the COVID-19	Chest CT images	Accuracy- 72%	The outcome of classification accuracy is comparatively low in the published year
Pavlova et al. [106], 2022	Improvements to deep convolutional neural networks	COVID-19 has been identified using a deep convolutional neural network with an upgraded architecture	Chest x-ray images	Accuracy- 96.3%	It is advised to design the model with more training images
Farooq et al. [107], 2020	ResNet	COVID-ResNet: A Deep Learning Framework for COVID19 Screening from Radiographs	Screening of radiography images	Accuracy-96.23%	It is expected to obtain more accuracy using the presented model
Awasthi et al. [108], 2021	Mini-COVIDNet	Mini-COVIDNet: An Effective Lightweight Deep Neural Network for COVID-19 Point-of-Care Ultrasound Detection	Lung ultrasound (US) imaging	Accuracy- 83.2%	It is suggested to apply Ct or X-rat images rather than ultrasound images with the concern of accuracy

**Table 4** (continued)

Author and year	Method used	Findings	Dataset	Outcome	Limitation
Hasan et al. [109], 2021	DenseNet	Application of DenseNet Convolutional Neural Networks for COVID-19 Prediction	CT Image	Accuracy- 92%	As new COVID-19 CT image datasets are regularly provided, updated COVID-19 imaging datasets may be examined to improve the efficiency of the DenseNet-121 method
Liu et al. [110], 2020	DenseNet with a two-dimensional sparse matrix profile	A Two-Dimensional Sparse Matrix Profile DenseNet for COVID-19 Diagnosis Using Chest CT Images	CT images	Accuracy- 78.07%	The need for classification accuracy was not reached to an effective measure
Xiao et al. [111], 2021	PAM-DenseNet	PAM-DenseNet, a deep convolutional neural network for computer-assisted COVID-19 diagnosis	Lung CT scan images	Accuracy- 94.29%	Due to imbalanced training data, training accuracy acquired is very less in the published year
Tang et al. [112], 2021	Ensemble deep learning-COVID	EDL-COVID stands for ensemble deep learning for COVID-19 case detection from chest X-ray images	Chest X-ray images	Accuracy- 93.3%	When compared to CXR photos of other classes, the number of COVID-19 CXR images in the COVIDx collection is still very low
Abdani et al. [113], 2020	Lightweight deep learning model	A deep learning model for detecting COVID-19	Chest X-ray images	Accuracy-94.6%	The accuracy level is not to the level due to the seriousness of COVID disease
Aminu et al. [114], 2021	Deep neural network	Deep neural networks and occlusion sensitivity maps for Covid-19 identification	CT scan images	Accuracy- 95.81%	Due to the severity of COVID illness, the accuracy level is not at the required level

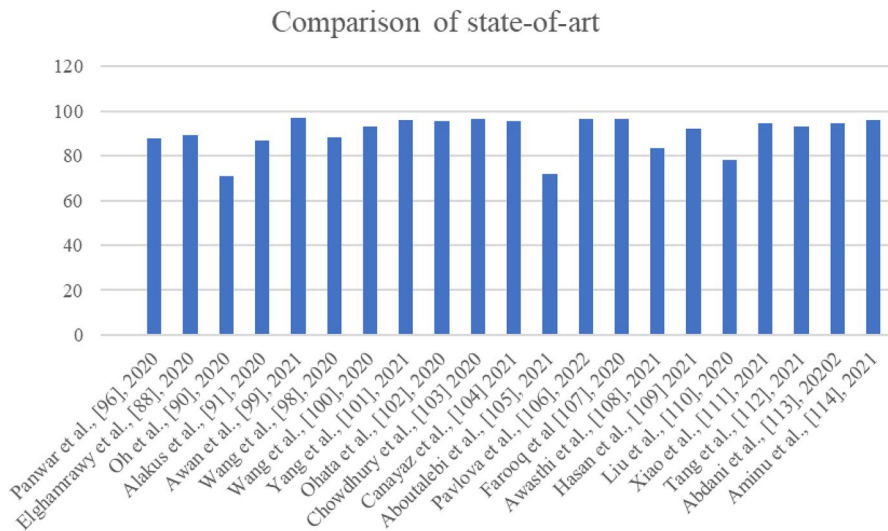


### 3.1 Challenges and Future Implications

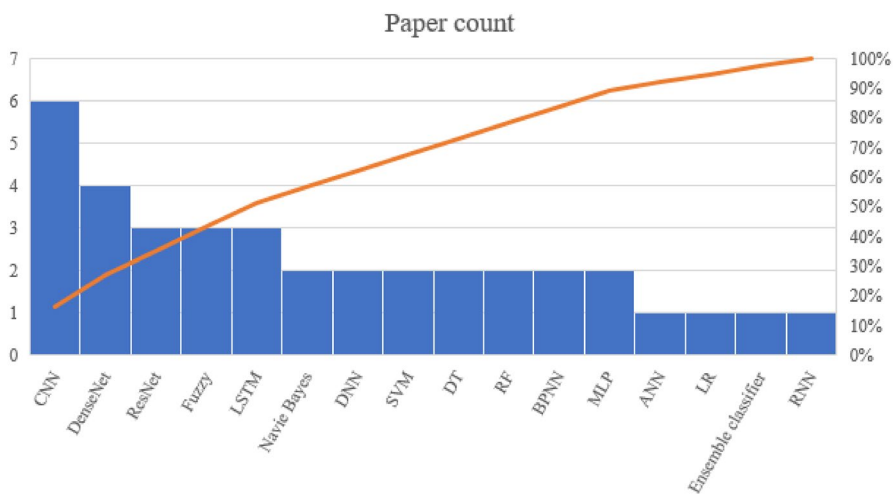
The invention and performance analysis of innovative data mining-based algorithms for COVID detection, prediction, and classification, as well as survey and review studies, have all contributed to the field of COVID diagnosis and prediction, as can be shown in [87–114]. We have provided an overview of cutting-edge strategies for combating the COVID-19 pandemic in this publication. Here, Big data applications for the COVID-19 illness, including as outbreak forecasting, tracking viral dissemination, diagnosis and treatment, and vaccine and medication discovery, have also been discussed. We have also spoken about the difficulties that must be solved if big data and AI are to effectively combat the COVID-19 epidemic. Finally, we have emphasized critical take aways and suggestions for the legal and scientific communities. On the basis of this analysis, it is determined that big data analytics has a significant impact and a promising future for research in the area of the COVID-19 illness diagnosis model. Big Data Analytics is used to extract useful information from COVID data, assisting COVID patients in determining the severity of the condition before it affects a person.

The novel coronavirus (COVID-19) pandemic has caused unprecedented disruption around the world, and it is clear that Artificial Intelligence (AI) will be critical in helping us to address this global challenge. AI can play a major role in identifying potential treatments for COVID-19, developing better diagnostic tests and predicting future outbreaks. It can also help healthcare organizations more effectively manage their resources by providing real-time data on patient health status and enabling them to make decisions quickly based on the latest available information.

- AI technologies such as machine learning algorithms have demonstrated great promise in aiding with disease diagnosis and prognosis. For example, AI systems are being used to identify patterns from medical images that might otherwise go unnoticed by human doctors or nurses when diagnosing patients with COVID-19 symptoms; this could potentially reduce misdiagnoses while speeding up treatment times significantly. In addition, predictive analytics models powered by AI are being used to forecast future outbreaks of the virus so governments can take steps early enough to contain them before they spread too far beyond control measures already put into place.
- Finally, natural language processing techniques enable automated chatbots which provide accurate answers about COVID prevention methods as well as other useful information about social distancing guidelines or local testing sites for those who need it most—all without requiring any direct contact between individuals during these uncertain times where physical contact should be avoided at all costs if possible! This ensures people get access to reliable advice while minimizing further risk of infection through face-to-face interactions whenever possible—an invaluable asset during these trying times indeed!
- Overall then it is clear just how powerful a tool Artificial Intelligence is proving itself out here: not only does its use aid healthcare professionals diagnose illnesses faster but also helps predict future outbreaks accurately so we may



**Fig. 4** Comparison of data mining technique in COVID diagnosis using big data analytics



**Fig. 5** Number of Published research articles taken for analysis

combat them more efficiently than ever before ultimately making our collective fight against COVID much easier going forward (Figs. 4, 5).

## 4 Conclusion

Our thorough analysis of the literature led us to the conclusion that most recent research either assesses the effectiveness of prevalent data mining-based COVID detection, prediction, and classification algorithms or provides succinct explanations of a small number of these tactics. To our knowledge, none of them, however, provides a thorough overview and comparison of the methods now in use and the pertinent difficult problems in big data analytics. Given the difficulties in the field of COVID identification, prediction, and classification based on big data analytics, we present a complete classification and comparison of existing methodologies utilising essential factors. Additionally, we provided a thorough state-of-the-art analysis on the creation of global systems for COVID diagnosis and prediction. We use big data analytics to break down machine learning and deep learning algorithms in order to accurately forecast when the COVID illness will manifest. The publications included in this study were acquired from a number of sources, including Wiley, IEEE Xplore, Nature, and ScienceDirect. To the best of our knowledge, imaging-based clinical diagnosis in the context of COVID diagnosis has not been directly addressed by the field of medical big data analytics. The relevant concerns are also examined, and a few unresolved problems are identified for more study in the area of big data analytics for the detection of communicable disease outbreaks. Additionally, it has been revealed that applying big data analytics in the case of the COVID pandemic can solve the current issue with illness diagnosis.

**Funding** Not applicable.

**Data Availability** Data sharing not applicable to this article as no datasets were generated.

## Declarations

**Conflict of Interest** The author declares there is no conflict of interest.

## References

1. Ranjan, J., Foropon, C.: Big data analytics in building the competitive intelligence of organizations. *Int. J. Inf. Manage.* **56**, 102231 (2021)
2. Mohamed, A., Najafabadi, M.K., Wah, Y.B., Zaman, E.A.K., Maskat, R.: The state of the art and taxonomy of big data analytics: view from new big data framework. *Artif. Intell. Rev.* **53**(2), 989–1037 (2020)
3. Mariani, M.M., Wamba, S.F.: Exploring how consumer goods companies innovate in the digital age: the role of big data analytics companies. *J. Bus. Res.* **121**, 338–352 (2020)
4. Mikalef, P., Krogstie, J.: Examining the interplay between big data analytics and contextual factors in driving process innovation capabilities. *Eur. J. Inf. Syst.* **29**(3), 260–287 (2020)
5. Holmlund, M., Van Vaerenbergh, Y., Ciuchita, R., Ravald, A., Sarantopoulos, P., Ordenes, F.V., Zaki, M.: Customer experience management in the age of big data analytics: a strategic framework. *J. Bus. Res.* **116**, 356–365 (2020)
6. Wong, Z.S., Zhou, J., Zhang, Q.: Artificial intelligence for infectious disease big data analytics. *Infect. Dis. Health* **24**(1), 44–48 (2019)

7. Manogaran, G., Shakeel, P.M., Baskar, S., Hsu, C.H., Kadry, S.N., Sundarasekar, R., Kumar, P.M., Muthu, B.A.: FDM: fuzzy-optimized data management technique for improving big data analytics. *IEEE Trans. Fuzzy Syst.* **29**(1), 177–185 (2020)
8. Li, W., Chai, Y., Khan, F., Jan, S.R.U., Verma, S., Menon, V.G., Li, X.: A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. *Mobile Netw. Appl.* **26**(1), 234–252 (2021)
9. Yasmin, M., Tatoglu, E., Kilic, H.S., Zaim, S., Delen, D.: Big data analytics capabilities and firm performance: an integrated MCDM approach. *J. Bus. Res.* **114**, 1–15 (2020)
10. Ghasemaghaei, M.: The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage. *Int. J. Inf. Manage.* **50**, 395–404 (2020)
11. Sousa, M.J., Pesqueira, A.M., Lemos, C., Sousa, M., Rocha, A.: Decision-making based on big data analytics for people management in healthcare organizations. *J. Med. Syst.* **43**(9), 1–10 (2019)
12. Aljumah, A.I., Nuseir, M.T., Alam, M.M.: Traditional marketing analytics, big data analytics and big data system quality and the success of new product development. *Business Process Manag. J.* **27**, 1108 (2021)
13. Peters, E., Klietnik, T., Musa, H., Durana, P.: Product decision-making information systems, real-time big data analytics, and deep learning-enabled smart process planning in sustainable industry 4.0. *J. Self-Governance Manag. Econ.* **8**(3), 16–22 (2020)
14. Mishra, S., Mishra, B.K., Tripathy, H.K. and Dutta, A.: Analysis of the role and scope of big data analytics with IoT in health care domain. In: *Handbook of data science approaches for biomedical engineering*, pp. 1–23. Academic Press, (2020)
15. Rehman, A., Naz, S. and Razzak, I.: Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Syst* 1–33. (2021)
16. Jia, Q., Guo, Y., Wang, G., Barnes, S.J.: Big data analytics in the fight against major public health incidents (Including COVID-19): a conceptual framework. *Int. J. Environ. Res. Public Health* **17**(17), 6161 (2020)
17. Ahn, P.D., Wickramasinghe, D.: Pushing the limits of accountability: big data analytics containing and controlling COVID-19 in South Korea. *Account. Audit. Account. J* (2021)
18. Awotunde, J.B., Ogundokun, R.O., Misra, S.: Cloud and IoMT-based big data analytics system during COVID-19 pandemic. In *Efficient data handling for massive internet of medical things* (pp. 181–201). Springer, Cham. (2021)
19. Verma, S. and Gazara, R.K.: Big data analytics for understanding and fighting COVID-19. In *Computational intelligence methods in COVID-19: Surveillance, prevention, prediction and diagnosis* (pp. 333–348). Springer, Singapore. (2021)
20. Shinde, P.P., Desai, V.P., Katkar, S.V., Oza, K.S., Kamat, R.K., Thakar, C.M.: Big data analytics for mask prominence in COVID pandemic. *Mater. Today* **51**, 2471–2475 (2022)
21. Zhou, H., Sun, G., Fu, S., Liu, J., Zhou, X., Zhou, J.: A big data mining approach of PSO-based BP neural network for financial risk management with IoT. *IEEE Access* **7**, 154035–154043 (2019)
22. Zhang, P., Shi, X., Khan, S.U.: QuantCloud: enabling big data complex event processing for quantitative finance through a data-driven execution. *IEEE Transact. Big Data* **5**(4), 564–575 (2018)
23. Wensheng, D.: Rural financial information service platform under smart financial environment. *IEEE Access* **8**, 199944–199952 (2020)
24. Zhou, H., Sun, G., Fu, S., Wang, L., Hu, J., Gao, Y.: Internet financial fraud detection based on a distributed big data approach with node2vec. *IEEE Access* **9**, 43378–43386 (2021)
25. Dos Anjos, J.C., Matteussi, K.J., De Souza, P.R., Grabher, G.J., Borges, G.A., Barbosa, J.L., Gonzalez, G.V., Leithardt, V.R., Geyer, C.F.: Data processing model to perform big data analytics in hybrid infrastructures. *IEEE Access* **8**, 170281–170294 (2020)
26. Yang, L., Yang, Y., Mgya, G.B., Zhang, B., Chen, L., Liu, H.: Novel fast networking approaches mining underlying structures from investment big data. *IEEE Transact Syst Man Cybern.* **51**(10), 6319–6329 (2020)
27. Ruan, J., Jiang, H., Yuan, J., Shi, Y., Zhu, Y., Chan, F.T., Rao, W.: Fuzzy correlation measurement algorithms for big data and application to exchange rates and stock prices. *IEEE Trans. Industr. Inf.* **16**(2), 1296–1309 (2019)
28. Sohagir, S., Wang, D., Pomeranets, A., Khoshgoftaar, T.M.: Big data: deep learning for financial sentiment analysis. *J. Big Data* **5**(1), 1–25 (2018)
29. Hassib, E.M., El-Desouky, A.I., El-Kenawy, E.S.M., El-Ghamrawy, S.M.: An imbalanced big data mining framework for improving optimization algorithms performance. *IEEE Access* **7**, 170774–170795 (2019)

30. Liu, B.: Text sentiment analysis based on CBOW model and deep learning in big data environment. *J. Ambient. Intell. Humaniz. Comput.* **11**(2), 451–458 (2020)
31. Zhai, G., Yang, Y., Wang, H., Du, S.: Multi-attention fusion modeling for sentiment analysis of educational big data. *Big Data Mining Anal.* **3**(4), 311–319 (2020)
32. Rodrigues, A.P. and Chiplunkar, N.N.: A new big data approach for topic classification and sentiment analysis of Twitter data. *Evolut. Intell.* 1–11 (2019)
33. Lau, R.Y.K., Zhang, W., Xu, W.: Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Prod. Oper. Manag.* **27**(10), 1775–1794 (2018)
34. Johnson, J.M., Khoshgoftaar, T.M.: The effects of data sampling with deep learning and highly imbalanced big data. *Inf. Syst. Front.* **22**(5), 1113–1131 (2020)
35. Juez-Gil, M., Arnaiz-González, Á., Rodríguez, J.J., García-Osorio, C.: Experimental evaluation of ensemble classifiers for imbalance in Big Data. *Appl. Soft Comput.* **108**, 107447 (2021)
36. Al, S., Dener, M.: STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment. *Comput. Secur.* **110**, 102435 (2021)
37. Juez-Gil, M., Arnaiz-González, Á., Rodríguez, J.J., López-Nozal, C., García-Osorio, C.: Approx-SMOTE: fast SMOTE for big data on apache spark. *Neurocomputing* **464**, 432–437 (2021)
38. Gupta, A., Lohani, M.C., Manchanda, M.: Financial fraud detection using naive bayes algorithm in highly imbalance data set. *J. Discrete Math. Sci. Cryptogr.* **24**(5), 1559–1572 (2021)
39. Kwon, J.M., Jung, M.S., Kim, K.H., Jo, Y.Y., Shin, J.H., Cho, Y.H., Lee, Y.J., Ban, J.H., Jeon, K.H., Lee, S.Y., Park, J.: Artificial intelligence for detecting electrolyte imbalance using electrocardiography. *Ann. Noninvasive Electrol. Cardiol.* **26**(3), e12839 (2021)
40. Sobanadevi, V. and Ravi, G.: Handling data imbalance using a heterogeneous bagging-based stacked ensemble (HBSE) for credit card fraud detection. In: *Intelligence in Big Data Technologies—Beyond the Hype*, pp. 517–525. Springer, Singapore. (2021)
41. Johnson, J.M. and Khoshgoftaar, T.M.: Thresholding strategies for deep learning with highly imbalanced big data. In: *Deep Learning Applications*, vol 2. Springer, Singapore, pp. 199–227 (2021)
42. Javaid, N., Jan, N., Javed, M.U.: An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. *J. Parallel Distributed Comput.* **153**, 44–52 (2021)
43. Arif, A., Javaid, N., Aldegheshem, A., Alrajeh, N.: Big data analytics for identifying electricity theft using machine learning approaches in microgrids for smart communities. *Concurr. Comput.* **33**(17), e6316 (2021)
44. Arif, A., Alghamdi, T.A., Khan, Z.A., Javaid, N.: Towards efficient energy utilization using big data analytics in smart cities for electricity theft detection. *Big Data Res.* **27**, 100285 (2022)
45. Hou, C., Wu, J., Cao, B., Fan, J.: A deep-learning prediction model for imbalanced time series data forecasting. *Big Data Mining and Analytics* **4**(4), 266–278 (2021)
46. Xia, D., Zhang, M., Yan, X., Bai, Y., Zheng, Y., Li, Y., Li, H.: A distributed WND-LSTM model on MapReduce for short-term traffic flow prediction. *Neural Comput. Appl.* **33**(7), 2393–2410 (2021)
47. Bawankule, K.L., Dewang, R.K. and Singh, A.K.: Historical data based approach to mitigate stragglers from the Reduce phase of MapReduce in a heterogeneous Hadoop cluster. *Cluster Comput.* 1–19 (2022)
48. Asif, M., Abbas, S., Khan, M.A., Fatima, A., Khan, M.A., Lee, S.W.: MapReduce based intelligent model for intrusion detection using machine learning technique. *J. King Saud Univ.-Comput. Inform. Sci.* **34**, 9723 (2021)
49. Wang, X., Wang, C., Bai, M., Ma, Q., Li, G.: HTD: heterogeneous throughput-driven task scheduling algorithm in MapReduce. *Distributed Parallel Databases* **40**(1), 135–163 (2022)
50. Pandey, V., Saini, P.: A heuristic method towards deadline-aware energy-efficient mapreduce scheduling problem in Hadoop YARN. *Clust. Comput.* **24**(2), 683–699 (2021)
51. Baruah, A.J., Baruah, S.: Data augmentation and Deep Neuro-Fuzzy network for student performance prediction with MapReduce framework. *Int. J. Autom. Comput.* **18**(6), 981–992 (2021)
52. Patan, R., Kallam, S., Gandomi, A.H., Hanne, T., Ramachandran, M., Gaussian relevance vector MapReduce-based annealed Glowworm optimization for big medical data scheduling. *J. Operat. Res. Soc.* 1–12. (2021)
53. Chawla, T., Singh, G., Pilli, E.S.: MuSe: a multi-level storage scheme for big RDF data using MapReduce. *J. Big Data* **8**(1), 1–26 (2021)

54. Narayana, S., Chandanapalli, S.B., Rao, M.S., Srinivas, K.: Ant cat swarm optimization-enabled deep recurrent neural network for big data classification based on map reduce framework. *Comput. J.* **65**, 3167 (2021)
55. Ramsingh, J., Bhuvanewari, V.: An efficient map reduce-based hybrid NBC-TFIDF algorithm to mine the public sentiment on diabetes mellitus—a big data approach. *J. King Saud University-Comput. Inform. Sci.* **33**(8), 1018–1029 (2021)
56. Roy, S., Bhattacharya, S., Omkar, S.N.: Automated Large-Scale Mapping of the Jahazpur Mineralised Belt by a MapReduce Model with an Integrated ELM method. *PFG J. Photogr. Remote Sens. Geoinform. Sci.* **90**(2), 191–209 (2022)
57. Pham, T.A., Dang, X.K., Vo, N.S.: Optimising Maritime Big Data by K-means Clustering with Mapreduce Model. In *International Conference on Industrial Networks and Intelligent Systems* (pp. 136–151). Springer, Cham, (2022)
58. Arunadevi, N., Thulasiraaman, V.: Cuckoo search augmented mapreduce for predictive scheduling with big stream data. *I. J. Sociotechnol. Knowledge Develop.* **14**(1), 1–18 (2022)
59. Kumar, D., Jha, V.K.: An improved query optimization process in big data using ACO-GA algorithm and HDFS map reduce technique. *Distributed Parallel Databases* **39**(1), 79–96 (2021)
60. Agarwal, G. and Om, H.: Parallel training models of deep belief network using MapReduce for the classifications of emotions. *Int. J. Syst. Assurance Eng. Manag.* 1–16. (2021)
61. Pang, Z., Wu, S., Huang, H., Hong, Z., Xie, Y.: AQUA+: Query Optimization for Hybrid Database-MapReduce System. *Knowl. Inf. Syst.* **63**(4), 905–938 (2021)
62. Maheswari, K., Ramakrishnan, M.: Kernelized Spectral Clustering based Conditional MapReduce function with big data. *Int. J. Comput. Appl.* **43**(7), 601–611 (2021)
63. Thanga Selvi, R., Muthulakshmi, I.: An optimal artificial neural network based big data application for heart disease diagnosis and classification model. *J. Ambient. Intell. Humaniz. Comput.* **12**(6), 6129–6139 (2021)
64. Ed-Daoudy, A. and Maalmi, K.: Real-time machine learning for early detection of heart disease using big data approach. In *2019 international conference on wireless technologies, embedded and intelligent systems (WITS)* (pp. 1–5). IEEE (2019)
65. Vaishali, G. and Kalaivani, V.: Big data analysis for heart disease detection system using map reduce technique. In *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)* (pp. 1–6). IEEE (2016)
66. Rastogi, R., Chaturvedi, D.K., Satya, S. and Arora, N.: Intelligent heart disease prediction on physical and mental parameters: a ML based IoT and big data application and analysis. In: *Machine Learning with Health Care Perspective*, pp. 199–236. Springer, Cham (2020)
67. Nayak, S., Gourisaria, M.K., Pandey, M. and Rautaray, S.S.: Comparative analysis of heart disease classification algorithms using big data analytical tool. In: *International Conference on Computer Networks and Inventive Communication Technologies*, pp. 582–588. Springer, Cham, (2019)
68. Nair, L.R., Shetty, S.D., Shetty, S.D.: Applying spark based machine learning model on streaming big data for health status prediction. *Comput. Electr. Eng.* **65**, 393–399 (2018)
69. Saluja, M.K., Agarwal, I., Rani, U. and Saxena, A.: Analysis of diabetes and heart disease in big data using MapReduce framework. In: *International Conference on Innovative Computing and Communications*, pp. 37–51. Springer, Singapore (2021)
70. Kılınc, D.: A spark-based big data analysis framework for real-time sentiment prediction on streaming data. *Software* **49**(9), 1352–1364 (2019)
71. Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., Benítez, J.M., Herrera, F.: Nearest Neighbor Classification for High-Speed Big Data Streams Using Spark. *IEEE Transact. Syst. Man Cybern.* **47**(10), 2727–2739 (2017). <https://doi.org/10.1109/TSMC.2017.2700889>
72. Park, K., Baek, C. and Peng, L.: A development of streaming big data analysis system using in-memory cluster computing framework: Spark. In: *Advanced Multimedia and Ubiquitous Engineering*, pp. 157–163. Springer, Singapore, (2016)
73. Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.A., Caelen, O., Mazzer, Y., Bontempi, G.: Scarff: a scalable framework for streaming credit card fraud detection with spark. *Inform. Fusion* **41**, 182–194 (2018)
74. Rathore, M.M., Son, H., Ahmad, A., Paul, A., Jeon, G.: Real-time big data stream processing using GPU with spark over hadoop ecosystem. *Int. J. Parallel Prog.* **46**(3), 630–646 (2018)
75. Zhou, B., Li, J., Wang, X., Gu, Y., Xu, L., Hu, Y., Zhu, L.: Online internet traffic monitoring system using spark streaming. *Big Data Mining Anal.* **1**(1), 47–56 (2018)

76. Xiao, W., Hu, J.: SWEclat: a frequent itemset mining algorithm over streaming data using Spark Streaming. *J. Supercomput.* **76**(10), 7619–7634 (2020)
77. Subramaniyan, S., Regan, R., Perumal, T. and Venkatachalam, K.: Semi-supervised machine learning algorithm for predicting diabetes using big data analytics. In *Business Intelligence for Enterprise Internet of Things*, pp. 139–149. Springer, Cham, (2020)
78. AlZubi, A.A.: Big data analytic diabetics using map reduce and classification techniques. *J. Supercomput.* **76**(6), 4328–4337 (2020)
79. Hatua, A., Subudhi, B.N., Veerakumar, T., Ghosh, A.: Early detection of diabetic retinopathy from big data in hadoop framework. *Displays* **70**, 102061 (2021)
80. Sivakumar, N.R. and Karim, F.K.D.: An IoT based big data framework using equidistant heuristic and duplex deep neural network for diabetic disease prediction. *J. Ambient Intell. Humanized Comput.* 1–11. (2021)
81. Mamatha Bai, B.G., Nalini, B.M., Majumdar, J.: Analysis and detection of diabetes using data mining techniques—a big data application in health care. In: *Emerging research in computing, information, communication and applications*, pp. 443–455. Springer, Singapore. (2019)
82. Sisodia, A., Jindal, R.: An effective model for healthcare to process chronic kidney disease using big data processing. *J. Ambient Intell. Humanized Comput.* 1–17 (2022)
83. Diez-Sanmartin, C., Sarasa-Cabezuelo, A., Belmonte, A.A.: The impact of artificial intelligence and big data on end-stage kidney disease treatments. *Expert Syst. Appl.* **180**, 115076 (2021)
84. Abdelaziz, A., Salama, A.S., Riad, A.M. and Mahmoud, A.N.: A machine learning model for predicting of chronic kidney disease based internet of things and cloud computing in smart cities. In: *Security in smart cities: models, applications, and challenges*, pp. 93–114. Springer, Cham, (2019)
85. Mansour, R.F., Escorcía-Gutiérrez, J., Gamarra, M., Díaz, V.G., Gupta, D., Kumar, S.: Artificial intelligence with big data analytics-based brain intracranial hemorrhage e-diagnosis using CT images. *Neural Comput. Appl.* 1–13. (2021)
86. HS, S.K. and Karibasappa, K.: An approach for brain tumour detection based on dual-tree complex Gabor wavelet transform and neural network using Hadoop big data analysis. *Multimedia Tools Appl.* 1–24 (2022)
87. Chew, A.W.Z., Pan, Y., Wang, Y., Zhang, L.: Hybrid deep learning of social media big data for predicting the evolution of COVID-19 transmission. *Knowl.-Based Syst.* **233**, 107417 (2021)
88. Elghamrawy, S.: An h 2 o's deep learning-inspired model based on big data analytics for coronavirus disease (covid-19) diagnosis. In *Big data analytics and artificial intelligence against COVID-19: Innovation Vision and Approach*, pp. 263–279. Springer, Cham (2020)
89. Jamshidi, M., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., Jamshidi, M., La Spada, L., Mirmozafari, M., Dehghani, M., Sabet, A.: Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. *Ieee Access* **8**, 109581–109595 (2020)
90. Oh, Y., Park, S., Ye, J.C.: Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans. Med. Imaging* **39**(8), 2688–2700 (2020)
91. Alakus, T.B., Turkoglu, I.: Comparison of deep learning approaches to predict COVID-19 infection. *Chaos Solitons Fractals* **140**, 110120 (2020)
92. Luo, Y., Xu, X.: Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. *Int. J. Hosp. Manag.* **94**, 102849 (2021)
93. Prasanth, S., Singh, U., Kumar, A., Tikkiwal, V.A., Chong, P.H.: Forecasting spread of COVID-19 using google trends: a hybrid GWO-deep learning approach. *Chaos Solitons Fractals* **142**, 110336 (2021)
94. Ramanathan, S., Ramasundaram, M.: Accurate computation: COVID-19 rRT-PCR positive test dataset using stages classification through textual big data mining with machine learning. *J. Supercomput.* **77**(7), 7074–7088 (2021)
95. Ghosh, T., Al Banna, M.H., Al Nahian, M.J., Taher, K.A., Kaiser, M.S. and Mahmud, M.: A hybrid deep learning model to predict the impact of COVID-19 on mental health form social media big data (2021)
96. Panwar, H., Gupta, P.K., Siddiqui, M.K., Morales-Menendez, R., Singh, V.: Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. *Chaos Solitons Fractals* **138**, 109944 (2020)
97. Kaur, H., Ahsaan, S.U., Alankar, B., Chang, V.: A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. *Inf. Syst. Front.* **23**(6), 1417–1429 (2021)
98. Wang, Y., Zeng, D.: Development of sports industry under the influence of COVID-19 epidemic situation based on big data. *J. Intell. Fuzzy Syst.* **39**(6), 8867–8875 (2020)

99. Awan, M.J., Bilal, M.H., Yasin, A., Nobanee, H., Khan, N.S., Zain, A.M.: Detection of COVID-19 in chest X-ray images: A big data enabled deep learning approach. *Int. J. Environ. Res. Public Health* **18**(19), 10147 (2021)
100. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Sci. Rep.* **10**(1), 1–12 (2020)
101. Yang, D., Martinez, C., Visuña, L., Khandhar, H., Bhatt, C., Carretero, J.: Detection and analysis of COVID-19 in medical images using deep learning techniques. *Sci. Rep.* **11**(1), 1–13 (2021)
102. Ohata, E.F., Bezerra, G.M., das Chagas, J.V.S., Neto, A.V.L., Albuquerque, A.B., de Albuquerque, V.H.C. and Reboucas Filho, P.P.: Automatic detection of COVID-19 infection using chest X-ray images through transfer learning. *IEEE/CAA Journal of Automatica Sinica*, 8(1), 239–248 (2020)
103. Chowdhury, N.K., Rahman, M., Kabir, M.A.: PDCOVIDNet: a parallel-dilated convolutional neural network architecture for detecting COVID-19 from chest X-ray images. *Health Inform. Sci. Syst.* **8**(1), 1–14 (2020)
104. Canayaz, M.: MH-COVIDNet: Diagnosis of COVID-19 using deep neural networks and meta-heuristic-based feature selection on X-ray images. *Biomed. Signal Process. Control* **64**, 102257 (2021)
105. Aboutaleb, H., Abbasi, S., Shafiee, M.J. and Wong, A.: COVID-Net CT-S: 3D convolutional neural network architectures for COVID-19 severity assessment using Chest CT Images. *arXiv preprint arXiv:2105.01284*. (2021)
106. Pavlova, M., Terhijan, N., Chung, A.G., Zhao, A., Surana, S., Aboutaleb, H., Gunraj, H., Sabri, A., Alaref, A. and Wong, A.: Covid-net cxr-2: an enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Front. Med.* **9** (2022)
107. Farooq, M. and Hafeez, A.: Covid-resnet: A deep learning framework for screening of covid19 from radiographs. *arXiv preprint arXiv:2003.14395*. (2020)
108. Awasthi, N., Dayal, A., Cenkaramaddi, L.R., Yalavarthy, P.K.: Mini-COVIDNet: efficient lightweight deep neural network for ultrasound based point-of-care detection of COVID-19. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **68**(6), 2023–2037 (2021)
109. Hasan, N., Bao, Y., Shawon, A., Huang, Y.: DenseNet convolutional neural networks application for predicting COVID-19 using CT image. *SN Comput. Sci.* **2**(5), 1–11 (2021)
110. Liu, Q., Leung, C.K., Hu, P.: A two-dimensional sparse matrix profile DenseNet for COVID-19 diagnosis using chest CT images. *IEEE Access* **8**, 213718–213728 (2020)
111. Xiao, B., Yang, Z., Qiu, X., Xiao, J., Wang, G., Zeng, W., Li, W., Nian, Y., Chen, W.: PAM-DenseNet: a deep convolutional neural network for computer-aided COVID-19 diagnosis. *IEEE Transact. Cybern.* **52**, 12163 (2021)
112. Tang, S., Wang, C., Nie, J., Kumar, N., Zhang, Y., Xiong, Z., Barnawi, A.: EDL-COVID: Ensemble deep learning for COVID-19 case detection from chest X-ray images. *IEEE Trans. Industr. Inf.* **17**(9), 6539–6549 (2021)
113. Abdani, S.R., Zulkifley, M.A. and Zulkifley, N.H.: A lightweight deep learning model for covid-19 detection. In: 2020 IEEE Symposium on Industrial Electronics & Applications (ISIEA) (pp. 1–5). IEEE (2020)
114. Aminu, M., Ahmad, N.A., Noor, M.H.M.: Covid-19 detection via deep neural network and occlusion sensitivity maps. *Alex. Eng. J.* **60**(5), 4829–4855 (2021)
115. Kumar, M.D., Ramana, K.: Cardiac Segmentation from MRI images using Recurrent & Residual Convolutional Neural Network based on SegNet and Level Set methods. *Annals of the Romanian Society for Cell Biology*, pp.1536–1545, (2021)
116. Kumar, M.D., Ramana, K.V.: Cardiovascular disease prognosis and severity analysis using hybrid heuristic methods. *Multimedia Tools Appl.* **80**(5), 7939–7965 (2021)
117. Mikkili, D.K.: Skin Cancer segmentation with the aid of multiclass dilated D-net framework. *Multimedia Tools Appl.* (2023). <https://doi.org/10.1007/s11042-023-14605-9>.
118. Rossetti, M., Pareschi, R., Stella, F., et al.: Integrating concepts and knowledge in large content networks. *New Gener. Comput.* **32**, 309–330 (2014). <https://doi.org/10.1007/s00354-014-0407-4>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



## Authors and Affiliations

**Nagamani Tenali<sup>1</sup> · Gatram Rama Mohan Babu<sup>2</sup>**

<sup>1</sup> Department of CSE, Dr.Y.S. Rajasekhara Reddy University College of Engineering & Technology, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, India

<sup>2</sup> Computer Science and Engineering (AI&ML), RVR & JC College of Engineering, Chowdavaram, Guntur, India