*International Journal of Modern Computation, Information, and Communication Technology*
*2023;6(2):51-58.*
***ISSN: 2581-5954***

# Big Data Analytics with Security Implementation Using Data Mining Algorithms

R. Kiruthika[1*], P. Devi[2], Vijayasarathi S[3], R M Asha[4]

[1]*Department of Chemistry, Sri Sairam Institute of Technology, Chennai, Tamil Nadu, India.*
[2]*PG & Research Department of Commerce, Thanthai Hans Roever College (Autonomous), Perambalur, Tamil Nadu, India.*
[3]*Department of Electrical and Electronics Engineering, Panimalar Institute of Technology, Chennai, Tamil Nadu, India.*
[4]*Department of Civil Engineering, Sri Sairam Institute of Technology, Chennai, Tamil Nadu, India*

[*]*Corresponding author: kiruthikar.chem@sairamit.edu.in*

**Abstract.** Big Data Analytics (BDA) is going to be discussed using data mining algorithms in this paper. Develop security in BDA fields for data security to achieve confidential storage. Big Data refers to data sets that are very vast, very complicated, and very dynamic and that have many independent sources. Big data is quickly implemented in all research and engineering fields, including physical, biological, and biomedical science, thanks to rapid developments in networking, data storage, and data gathering capacities. From a data mining point of view, this study gives a theorem characterizing the aspects of the Big Data revolution and suggests a Big Data processing model. This data-driven paradigm considers security and privacy concerns, data mining and analysis, and the modeling of user interests. Here, takes a look at a few of the knottier aspects of big data and the data-driven approach, and calculate the precisions, recalls, and measures, which are 90% of different data mining algorithms.

**Keywords:** Data mining, Data abstraction, Data management, Big Data, Data mining algorithms.

## INTRODUCTION

Hybridization of cloud computing, data mining, and large amounts of available internet data is a topic of discussion in this study. Data visualization is a technique used in data mining for the purpose of studying and making sense of massive data sets. Convention and algorithms in computing have been researched for their impact on data storage and transfer needs. The perspective of data analysis is often investigated in studies of various techniques for massive data storage. Methodological refinement and problem statements have been addressed using this vocabulary and these considerations. This will aid the pursuit of fresh information and the examination of computing capability. About fifteen articles compared the ways in which big data is used. This work investigates Big Data Mining strategies for the cloud and discusses cloud-friendly issues and computational strategies for advancing Big Data Mining in the cloud [1]. The Internet of Things (IoT) is a new paradigm that may provide incredible insights via data mining and analysis. The IoT is a vision for a future in which every internet-enabled device—from mobile phones to cars to public buildings to domestic appliances—serves as a data source. Even in the present day, a wide variety of electronic gadgets, such as watches, emergency alarms, parking doors, and numerous appliances, may be connected to IoT networks and operated from afar. The performances of IoT systems and their ability to store, process, and analyze data provide unique problems that may be handled with the help of big data analytics and data mining techniques. Extensive research with big data for the IoT may provide a wealth of data that can be mined for insights. Considering this, this study presents a comprehensive literature review on the uses of big data analytics and data mining techniques in the IoT [2].

The purpose of this overview is to determine which areas of study need more exploration in forthcoming publications. To do so, I looked at all the literature on IoT big data and IoT data mining published between 2010 and 2021 (a total of 60 papers). The topics covered in these articles may be broken down into four broad categories: architectures/platforms, frameworks, applications, and securities. This study summarizes the approach used in big data analysis and data mining using the IoT in these four areas to highlight the most fruitful lines of inquiry for

*Received: 09.08.2023 Revised: 28.09.2023 Accepted: 13.10.2023*

future studies [3]. Although public databases like Surveillance, Epidemiology, and End Result (SEERs), the National Health and Nutrition's Examinations Surveys (NHANESs), The Cancers Genome Atlas (TCGAs), and the Medical Information's Marts for Intensive Cares (MIMICs) have produced many high-quality studies, the data in these databases are often characterized by a higher degree of dimensional heterogeneities, timeliness, scarcities, irregularities, and other characteristic that prevent their full value from being realized. Since data-mining technology has shown such promising results in assessing patient risk and aiding clinical decision-making in developing disease-predictions models, it has emerged as a cutting-edge area of study in the medical sciences.

Since there are so many enormous public medical databases, data mining lends itself particularly well to clinical big-data research. This article provided an accessible introduction to the primary public medical database and outlined the processes, activities, and models involved in data mining. Also, data-mining techniques and their uses are detailed. The purpose of this effort was to provide clinical researchers with a completer and more instinctive grasp of how to use data-mining technologies to clinical big data for the benefit of medical practitioners and their patients [4]. This paper joins data mining technologies with up-to-date agricultural data to address the issues of stale data and missing data in the agricultural data collection and sorting processes inherent to the Internet era of data mining. To further enhance agricultural data mining and statistical algorithms, these work implements novel approaches to time representations and time series measurements. In addition, practical mining techniques for learning from data analysis are discovered in this work. Finally, this study integrates the real requirements for building data mining and statistical analytics models of precision and intelligence agricultures based on big data analysis and proposes experiments to validate the model's performances. The findings demonstrate the usefulness of the model developed in this work.

5G will be a game-changer due to the massive expansion of several sectors. Supporting transportation infrastructure, 5G networks are all about fast speeds, low latency, and wide coverage. The 5G network has been a benefit to the IoT business since it improves device performance and dependability. Since the 5G spectrum increases the frequency on which digital cellular technologies carry data, they work as a dominant to the IoT by providing the turbo network it requires to increase capacity and improve connection. The need to automate and investigate all the appliances and sensors, as well as keep up with them, gives birth to the IoT [5]. From troubleshooting complicated installations to managing network traffic, this chapter covers it all. Information extraction is handled by Machine Learning (ML) and Artificial Intelligence (AI), whereas data mining is a built-in process in 5G IoT that simplifies decision-making and problem-solving. The importance of big data analytics, which is discussed in this chapter, is highlighted in the context of mobile networks. Increased productivity, shorter decision cycles, and robust real-time analytics are just a few of the benefits of 5G's successful technology integration. This paper's primary goal is to fill in the theoretical gaps around the topic of big data mining for clients' insight by describing the limitations of the temporal approaches to big data analyses found in existing scientific literature. Two types of study are used in this article. The first approach is to look for topics related to big data mining for consumer insights by conducting a systematic search of bibliographic repositories. Four stages of this procedure have been completed [6].

Bibliographically verifying the findings is the second research strategy. The verification process included doing a Scopus search using the predetermined keywords and then analyzing the results for any discernible patterns. This study's main contributions are an organized body of knowledge on the roles of the latest BDAs, primarily big data mining, in comprehending customers' behaviors; an indication of the importance of the temporal dimensions of clients' behaviors; and the identification of interesting research gaps: the mining of temporal big data for a comprehensive picture of customer [7]. The problem statement is discussed below. There are many security issues in BDA storage, processing, and analysis. To achieve security in these areas, use data mining algorithms to avoid security risks. Data mining methods would be very helpful in many sectors to achieve the security and privacy of data.

The following are the contributions.

- BDA has made a significant impact on database storage
- Using data mining algorithms, BDA helps to identify the fraudulent attacks
- Organizations use these data mining algorithms to avoid security breaches in banking, insurance, and cyber security sectors. The following section will be a literature survey discussed in section 2, and the proposed

system using data mining algorithms for BDA will be discussed in section 3. The Result and discussion are discussed for the given dataset to improve the precision, recall, and F measure in section 4. Finally, the conclusion provides the overall performance of the BDA and future work.

## LITERATURE SURVEY

All industries have embraced digitalization and the progress of information technology, particularly AI. Simultaneously, data has expanded largely thanks to digital traces or other technological information systems. Opportunities for BDA in all sectors, including education, have never been greater than they are now. The purposes of this study are to provide a comprehensive overview of the literature on the uses of BDA in the fields of education and to outline potential future directions for investigation. I chose and grouped the literature according to the data kinds, techniques, data analytics, and learning analytic software used, using Kitchenham's approach. In the context of massive open online courses (MOOCs) and learning management systems (LMSs), the findings suggest that big data learning analytics research is typically aimed at bettering the learning experience, analyzing learner behavior for students profiling, improving student retention, and evaluating student's feedback [8].

There are a number of promising avenues of research for this area, including the creation of a large, publicly available dataset, complete with data pre-process and the resolution of the imbalanced dataset problem, and the implementation of processes mining for learning log activities in order to glean knowledge and insight from online behavior, both from the perspectives of the student and the instructor and Developing an automated system that takes advantage of big data to conduct analytical learning at the descriptive, predictive, and prescriptive levels. In conclusion, using big data for learning analytics and educational data mining seems to be a promising open research subject in the field of education [9]. Medical imaging serves crucial roles in the diagnosis of illness. The amount of local storage space and transfer bandwidths needed by remote medical devices to aid patient diagnosis and treatment is strongly related to how well it compresses data. Lossless compression and similarity are two amazing features of medical imaging. The crux of compression is figuring out how to make use of these two features to decrease the amount of data required to describe a picture. In this work, big data mining is used to create an image catalog. To break down a picture into its constituent parts. To accurately portray the essential structure of multi-component medical pictures, suggest a soft compression approach. Results show that the created soft compression technique can exceed the common benchmark PNG and JPEG2000 in terms of compression ratio, and a comprehensive representation framework for image compressions is also proposed [10].

The prevalence of diabetes has increased dramatically over the world. Chronic tissue damage may result from prolonged exposure to hyperglycemia. So, it's important to detect diabetes early and treat it. In this research, an algorithm to predict diabetes risk by combining information from several clinical examinations is developed. Data on 1,507,563 persons in Luzhou City, China, who were either healthy or diagnosed with diabetes, as well as 387,076 people with diabetes who were followed up between 2011 and 2017, is gathered. The demographics, vital signs, and laboratory data of a patient's medical examination were subjected to statistical analysis. An extreme gradient boosting (XGBoost)--based model was created to differentiate between persons with and without diabetes, with an AUC of 0.8768 being achieved. Further, the diabetes risk scorecards were built based on logistics regressions, which could assess human health to enhance the model's conveniences and adaptabilities in clinical and real-world settings. Finally, statistical analysis to extract the more important characteristics affecting the patient's ability to keep their disease under control from their follow-up records is used [11].

Risk variables associated with cyberbullying in Korea were analyzed using a decision tree analysis based on big social data. The research analyzed 103,212 buzzes from 227 online channels such as news websites, blogs, online groups, social network services, and online bulletin boards to determine what factors contribute to cyberbullying 25.0 was used to classify the various forms of cyberbullying by opinions-mining and decisions tree analysis. Based on the data, it was determined that there was a 44% prevalence rate for all forms of cyberbullying in Korea, with 32.3% of the population being bullied, 6.4% of the bullies, and 5.5% of the spectators.

The findings showed that the inclination for dominance was the second most influential component in predicting the kinds of risk factors, with the impulsive factors being the most influential element overall. Spectators were most influenced by the impulsive component, whereas online offenders were most influenced by the inclination for dominance factor. Because many onlookers tend to intensify impulsive cyberbullying actions, it is vital to establish a program to lessen the impulses that were triggered by both victims and offenders [12].

Using a hybrid approach, including the generalized additive models and the segmented regression models, this research quantifies the effects of climate change on hiking in 100 cities throughout China. The findings show that environmental factors like temperatures, relative humidity, and sunlight durations have nonlinear and threshold impacts on the number of people who go hiking. Over 90% of the cities analyzed in simulation research had negative impacts on hiking due to climate change. Under RCP 4.5, the time spent on hikes would decrease by 7.17–7.39 percent in 2050 and 7.16–7.57 percent in 2080. The scenario under RCP 8.5 is considerably drier. Advocate for this method to be used in other locations that have access to similar data.

The steel sector has been a pioneer in the use of big data. The manufacturing of iron requires the cooperation of many different industries, all of which always create a tremendous amount of data. To properly store and make full use of the iron production data, it is necessary to construct big data platforms. To maximize output while minimizing energy use without sacrificing quality or durability, a method to evaluate its current state and anticipate its future performance. With its foundation in a big data platform and its incorporation of a factor analysis approach, the evaluative system can both identify and extract the latent common components in the production index when 19 state parameters are considered. It can then proceed to compute the whole status index [13]. The AdaBoost model used by the prediction system allows for precise forecasting of the status indexes three hours in advance. Based on the evaluations, it seems that the recommended status indexes closely match the actual state throughout the chosen time. Factor analysis also confirms the degree of congruence between the status index over time periods and the real scenario. The assessment and prediction system shows great accuracy in the present production environment, but owing to the long-term changes in production, it may still require calibration and updates on a regular basis. When used properly, the online comprehensive assessment and forecast system may greatly aid operators in improving operations and keeping them stable.

The medical industry has benefited quickly from the proliferation of cutting-edge technologies. A major stage of risk prediction, the current epidemic of coronavirus diseases 2019 () has quickly become an epidemic to discover early actions from suspected patients. Intense effort is required to develop a monitoring infrastructure that can track down COVID-19. Reverse transcriptions-polymerase chains reactions (rRT-PCR) by viral extensions is currently the gold standard for confirming COVID-19 infection, despite being identified from deficiency of long reversals times to produce Results in 2-4 h of coronas with a requirement of a certified laboratory [14]. The suggested system utilizes ML algorithms and the textual data mining technique to categorize the clinical report into four distinct types. A sophisticated information retrieval method called term frequencies-inverses document frequencies (TF/IDF) was used by the ensemble ML classifier algorithm to extract features from the corona dataset. Human coronaviruses (HCoVs) NL63, OC43, HKU1, and 229E cause mild respiratory diseases that are globally pandemic; zoometric Middle East respiratory syndromes coronaviruses (MERS-CoV); and severe acute respiratory syndromes coronaviruses (SARS-CoV) cause higher casualty rates [15].

## PROPOSED SYSTEM

Even with Big Data, there remain problems to solve, such as how to collect, transport, store, clean, analyze, filter, search, share, protect, and visualize information. Even still, one of the primary challenges in this field is the storage and retrieval of huge amounts of data. When dealing with Big Data issues, it's important to strike a balance between the data's scalability, availability, performance, and security needs. Big data aims to examine complex and changing relationships among data by using huge volumes of diverse, independent sources with distributed and decentralized governance. Due to these factors, mining Big Data for insights is very difficult. One simplistic analogy is many blind individuals attempting to assess the size of an elephant, which represents Big Data. Each blind man's task is to piece together an image of the elephants using the information he receives. Because people could only see so far beyond their immediate surroundings. Big data analysis in this situation is analogous to a group of blind guys combining their senses to create the most accurate image possible of an elephant's true gesture. It's not as easy as just asking how the blind men feel about the elephant. The next step is to have an expert design a single image that represents everyone's perspective, even though they may not all understand each other and may have privacy concerns over the messages they discuss. Many industries may gain a competitive edge and see a boost in value through smarter data mining. The three Vs are the cornerstone qualities of Big Data. Figure 1 shows the system architecture of the proposed system.

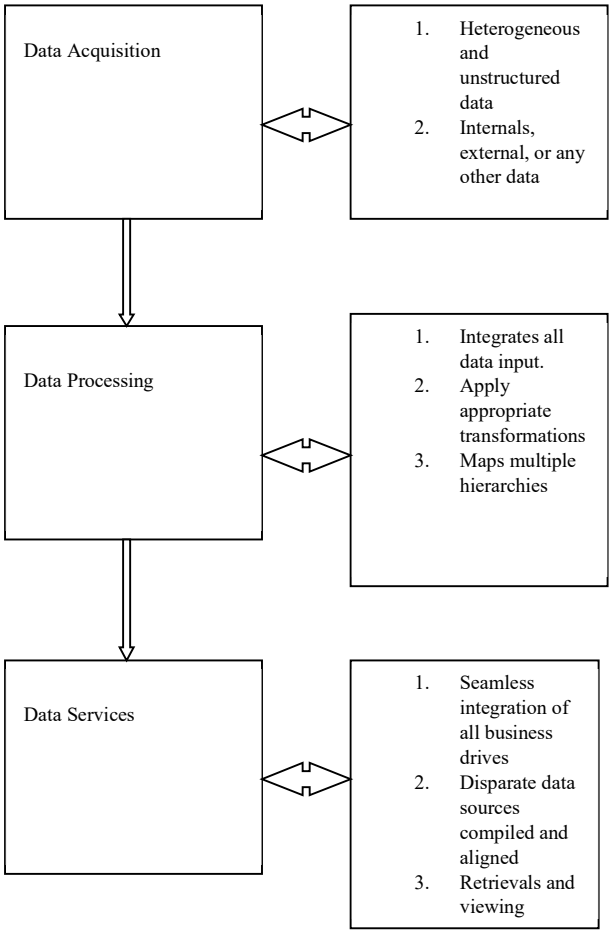


**FIGURE 1.** System architecture of the proposed system

The three Vs are the fast expansion and evolution of data, the wide range of data types and formats, and the massive volumes of data created every second. Different types of research groups have been working hard to create new, quick, dynamic, and user-friendly Big Data technology. The scattered nature of today's networks and the complexity of Big Data mean that conventional search techniques no longer suffice. Businesses require the ability to query massive amounts of unstructured and structured data quickly and efficiently. Because of this need, advanced search engines with scalable search and indexing capabilities have been created. Obtaining the correct data through transformations in Big Data is also crucial. The possibilities Big Data offers for data-driven scientific and technological breakthroughs are enormous. The potential of Big Data to boost creativity, rivalry, and output is promising. For optimal speed, trustworthy findings, and scalability, Big Data modeling and mining necessitate the use of very sophisticated technology and techniques

## RESULTS AND DISCUSSIONS

There is still a need for investigation into several topics that might improve the capabilities and features of Big Data applications. The proposed study is centered on unified architecture for collecting, transmitting, storing, and analyzing data on a scale. After that, provide services for retrieving and processing real-time data sets, such as those from the stock market, financial sensors, and the medical field. Efficiency, processing time, flexibility, and scalability are only a few of the aspects of performance that must be analyzed across the work's many abstraction layers. Let's try calculating precision, recall, and F-measure for the below models that classified 100 tumors as malignant (the positive class) or benign (the negative class) medical dataset shown in Table 1.

**TABLE 1.** Medical Dataset

| | | |
|---|---|---|
| True Positive (TP):<br>Reality: Malignante<br>DM model predicted: Malignant<br>Number of TP results: 1 | False Positive (FP):<br>Reality: Benign<br>DM model predicted: Malignant<br>Number of FP results: 1 | True Positive (TP):<br>Reality: Malignante<br>DM model predicted: Malignant<br>Number of TP results: 1 |
| False Negative (FN):<br>Reality: Malignant<br>DM model predicted: Benign<br>Number of FN results: 8 | True Negative (TN):<br>Reality: Benign<br>DM model predicted: Benign<br>Number of TN results: 90 | False Negative (FN):<br>Reality: Malignant<br>DM model predicted: Benign<br>Number of FN results: 8 |

A unified Big Data platform for real-time applications is at the heart of the proposed study. The three tiers of abstraction are as follows: The components of data collection, data transmissions, and data pre-processing are all part of the larger process known as "data acquisition." Data processing, also known as information integration and data storage, is the act of transforming raw information into usable information. Services for retrieving and using data are provided by this category. This serves as an interface, allowing the users to quickly and easily compile data from many sources. Figure 2 shows the frequency of the top 5 data mining techniques, and Figure 3 shows the performance of different Data mining algorithms.

$$Frequency = \frac{Number\ of\ times\ the\ data\ occured}{Length\ of\ the\ time} \quad (1)$$

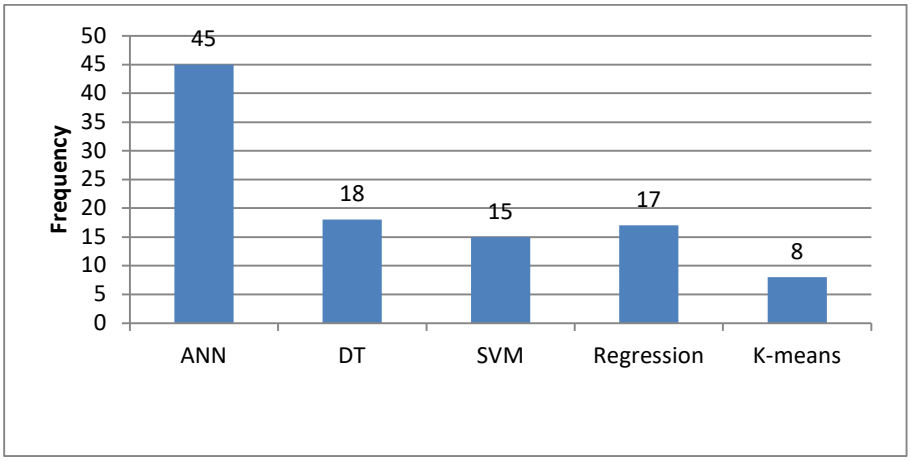


**FIGURE 2.** Frequency of top 5 data mining techniques

$$Precision = \frac{True\ positive}{total\ Predicted\ positive} \quad (2)$$

$$F - measure = \frac{Precision * Recall}{Precision + Reca} \quad (3)$$

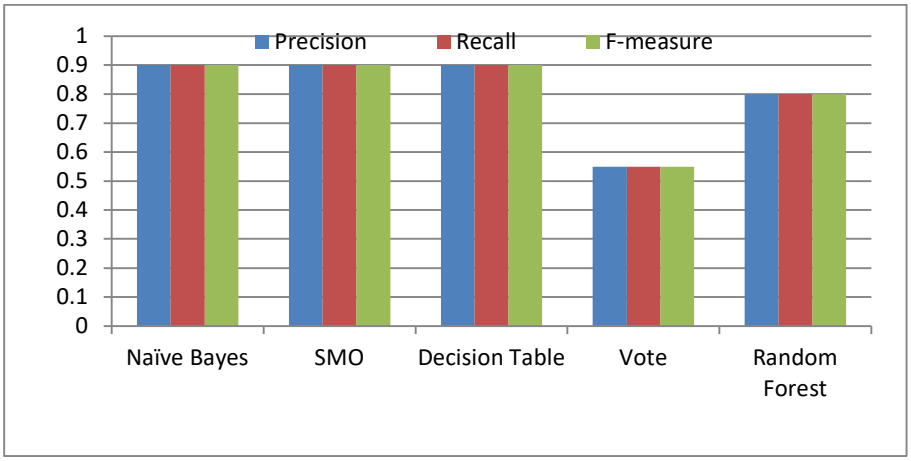


**FIGURE 3.** Performance of different algorithms

A lot of work has been done to figure out how to make use of big-time series databases as storage and processing power have improved. Real numbers representing values at discrete times make up a time series. Stock prices, sensor data, picture borders, and biomedical data are all examples of time series data. The infrastructure needed to acquire Big Data should provide low, predictable latencies when capturing data and when executing short, simple queries; it must be able to handle very high transaction volumes, often in distributed environments; and it must support flexible, dynamic data structure. In comparison to finding, recognizing, comprehending, and citing data, data processing presents a far greater challenge. All of this must be handled mechanically for large-scale analysis to work. To achieve this goal, discrepancies in data structures and semantics must be stated in forms that can be understood by computers and "robotically" resolved. Data integration, mapping, and transformations have a lot of research behind them. There is still a lot of work to be done before automated, error-free difference resolution can be achieved. Big data querying and mining techniques are very unlike the tried-and-true methods of analyzing small samples of data. The integration and ease of access to data provided by data services are invaluable. The data itself, its reliability, its semantics, and the intelligence of its querying operations may all be enhanced by data mining.

## CONCLUSIONS

Initiated by government funding agencies and propelled by real-world applications and important industry players, big data management and mining have proven to be arduous but ultimately rewarding endeavors. Big Data's major qualities include autonomy via dispersed and decentralized control, complexity through always changing data and knowledge linkages, and scale, as implied by the term's literal focus on data quantities. Large Data's combination of features suggests that it needs a "big mind" to extract the most value from data consolidation. Numerous difficulties at the levels of data, models, and systems in the investigation of Big Data are examined. To fully use the potential of Big Data, high-performance computing systems are needed to facilitate Big Data mining. Complex circumstances, such as missing or unknown values, are typically present in data due to the independent information sources and the diverse data-gathering contexts and observed frequency, F-measure, and precision of the proposed system.

## REFERENCES

[1]. Z.S. Agreed, S. R. Zeebaree, M.M. Sadeeq, S.F. Kak, H.S. Yahia, M.R. Mahmood, and I.M. Ibrahim, 2021, “Comprehensive survey of big data mining approaches in cloud systems,” *Qubahan Academic Journal*, **1(2)**, pp. 29-38.

[2]. Y. Zhong, L. Chen, C. Dan, and A. Rezaeipanah, 2022, “A systematic survey of data mining and big data analysis in the Internet of things,” *The Journal of Supercomputing*, **78(17)**, pp. 18405-18453.

[3]. W.T. Wu, Y.J. Li, A.Z. Feng, L. Li, T. Huang, A.D. Xu, and J. Lyu, 2021, “Data mining in clinical big data: the frequently used databases, steps, and methodological models,” *Military Medical Research*, **8**, pp. 1-2.

[4]. Z. Rao, and J. Yuan, 2021, "Data mining and statistics issues of precision and intelligent agriculture based on big data analysis," *Acta Agriculture Scandinavica, Section B—Soil and Plant Science*, **7(9)**, pp. 870-883.

[5]. P.K. Aggarwal, P. Jain, J. Mehta, R. Garg, K. Makar, and P. Chaudhary, 2021, "Machine learning, data mining, and big data analytics for 5G-enabled IoT," *Blockchain for 5G-Enabled IoT: The new wave for Industrial Automation*, pp. 351-375.

[6]. M. Mach-Król, and B. Hadasik, 2021, "On a certain research gap in big data mining for customer insights," *Applied Sciences*, **11(15)**, pp. 1-6.

[7]. A. Yunita, H.B. Santoso, and Z.A. Hasibuan, 2021, "Research review on big data usage for learning analytics and educational data mining: A way forward to develop an intelligent automation system," *In Journal of Physics: Conference Series, IOP Publishing*, **1898(1)**, pp. 1-7.

[8]. G. Xin, and P. Fan, 2021, "A lossless compression method for multi-component medical images based on big data mining," *Scientific Reports*, **11(1)**, pp. 1-9.

[9]. H. Yang, Y. Luo, X. Ren, M. Wu, X. He, B. Peng, K. Deng, D. Yan, H. Tang, and H. Lin, 2021, "Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators," *Information Fusion*, **75**, pp. 140-149.

[10]. T.M. Song, and J. Song, 2021, "Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data," *Telematics and Informatics*, **58**, pp. 1-7.

[11]. J. Liu, L. Yang, H. Zhou, and S. Wang, 2021, "Impact of climate change on hiking: quantitative evidence through big data mining," *Current issues in tourism*, **24(21)**, pp. 3040-3056.

[12]. H. Li, X. Bu, X. Liu, X. Li, H. Li, F. Liu, and Q. Lyu, 2021, "Evaluation and prediction of blast furnace status based on big data platform of ironmaking and data mining," *ISIJ International*, **61**(1), pp. 108-118.

[13]. S. Ramanathan, and M. Ramasundaram, 2021, "Accurate computation: COVID-19 rRT-PCR positive test dataset using stages classification through textual big data mining with machine learning," *The Journal of Supercomputing*, **77**, pp. 7074-7088.

[14]. C. Pieroni, M. Giannotti, B.B. Alves, and R. Arbex, 2021, "Big data for big issues: Revealing travel patterns of low-income population based on smart card data mining in a global south unequal city," *Journal of Transport Geography*, **96**, pp. 1-5.

[15]. S. Liang, "Research on the method and application of mapreduce in mobile track big data mining,"2021, *Recent Advances in Electrical and Electronic Engineering*, **14(1)**, pp. 20-28.