ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Efficient Breast Cancer Dataset Analysis Based on Adaptive Classifiers

Thikra Ali Kareem[1], Muzhir Shaban Al-Ani[2], Salwa Mohammed Nejres[3]

[1]Imam Ja'afar Al-Sadiq University, Salah Al-Din, Dijail, [2]Department of Information Technology, College of Science and Technology, University of Human Development, Sulaymaniyah, KRG, Iraq, [3]Ministry of Higher Education and Scientific Research, Baghdad, Iraq.

ABSTRACT

Many algorithms have been used to diagnose diseases, with some demonstrating good performance while others have not met expectations. Making correct decisions with the minimal possible errors is of the highest priority when diagnosing diseases. Breast cancer, being a prevalent and widespread disease, emphasizes the importance of early detection. Accurate decision-making regarding breast cancer is crucial for early treatment and achieving favorable outcomes. The percentage split evaluation approach was employed, comparing performance metrics such as precision, recall, and f1-score. Kernel Naïve Bayes achieved 100% precision in the percentage split method for breast cancer, while the Coarse Gaussian support vector machines achieved 97.2% precision in classifying breast cancer in 4-fold cross-validation.

Index Terms: Breast cancer, Adaptive classifiers, Performance measures, Percentage split evaluation technique, Healthcare analysis.

1. INTRODUCTION

The development of the Internet and its containment of a vast amount of data, coupled with the emergence of large data volumes generated by social media, has made data today exceptionally massive. This necessitates new approaches for its transfer, processing, and analysis, collectively known as big data. This data may be unstructured, semi-structured, or structured, benefiting various sectors such as societal, business, industry, agriculture, education, and healthcare [1], [2].

The internet of things (IoT) contains a huge number of sensors connecting to other systems and devices, producing extensive data that requires novel analytical methods, particularly wireless sensor networks, as the data source [3], [4].

Given the immense volume of data, traditional database systems are inadequate for storing, handling, and analyzing such quantities. The concept of big data involves large-scale processes for identifying and interpreting information into new insights. While the term "big data" has long been in use, its prominence surged following the rise of social media [5], [6]. The exponential growth of data worldwide, characterized by vast volume, high speed, and diverse types, demands an infrastructure capable of simultaneous processing and storage. Cloud computing offers on-demand computing resources that can be quickly configured, provisioned, and released as per users' needs, making it accessible to individuals and organizations alike [7], [8]

In data analysis, starting from processing, cleaning, and filtering, correct data retrieval is essential to perform the necessary analysis. Analyzing and calculating metrics such as maximum, minimum, and standard deviation using relevant tools are

Access this article online

DOI:10.21928/uhdjst.v8n1y2024.pp122-128 | E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Department of Information Technology, College of Science and Technology, University of Human Development, Sulaymaniyah, KRG, Iraq. Email: muzhir.al-ani@uhd.edu.iq

Received: 06-02-2024 Accepted: 06-04-2024 Published: 25-04-2024

crucial for obtaining accurate results. Optimal analysis follows proper steps to ensure accuracy and reliability [9], [10]. Smart cities can utilize distributed stream processing frameworks for real-time data processing, a practical application in addition to their current IoT adoption. Choosing a suitable framework for smart city data analytics requires a comprehensive understanding of target application features [11], [12].

Big data analysis employs various analytical methods depending on the data's nature, aiming to extract insights. These methods include statistical analysis, correlation analysis, and regression analysis [13], [14]. The aim of the proposed approach is to develop an effective algorithm for disease prediction by employing classification approaches to address health-care issues during diagnosis and achieve efficient performance results.

2. LITERATURE REVIEW

Babar *et al.* proposed a health-care architecture based on energy harvesting analysis of health monitoring devices. The proposed architecture consists of three layers. Consistent datasets were verified on Hadoop server to validate the proposed design based on the threshold value. The goal of this research is to make smart decisions and deal with events. The analysis shows that the proposed design has great potential in the field of smart health [15].

Singh and Yassine utilized IoT and big data analytics to create energy management strategies to manage home energy effectively and efficiently. They proposed a unified architecture that enables creative activities to process massive amounts of granular energy use data in close to real-time. The complexity and resource requirements of data processing, storage, and classification analysis in close to real-time are addressed by proposing an IoT big data analysis system that uses fog computing [16].

Shah *et al.* proposed a new design and philosophy for the disaster-resistant smart city. Together, the Hadoop ecosystem and Spark form a powerful system environment that allows for real-time and offline analysis. System efficiency is evaluated in terms of processing time and throughput. The aim of this research is to add to the body of knowledge and guide future research on the design and implementation of disaster-resilient smart cities based on this system and the IoT. This strategy can lead to immediate and effective situational awareness, which can help mitigate the effects of the disaster [17].

Syed *et al.* proposed a new smart healthcare framework to monitor the physical activity of older individuals using the Internet of Medical Things and intelligent machine learning algorithms for quick analysis, decision-making and better treatment suggestions. Hadoop MapReduce algorithms are used to handle massive amounts of data in parallel. The aim of this study is to predict the physical activity of respondents to help them live a healthier lifestyle. This study provides an excellent option for detecting physical activity and monitoring the health of elderly people remotely [18].

Li introduced a fog-assisted IoT-based intelligent and real-time health-care information processing system. In this system, minimal latency and large amounts of data generated by IoT sensors are offloaded to the fog cloud for data analysis and processing. The data then are processed and stored in a central cloud system using Hadoop and Apache Spark in order to process and analyze. The proposed compression strategy results in a 60% reduction in the amount of data in this system, in addition, for real-time data analytics; it offers a fog-powered approach with a big data environment [19].

Hadi *et al.* presented a multidisciplinary approach to e-health care, priority, big data analytics, and radio resource optimization in a multi-tier 5G network. They used a combined system including three machine learning algorithms (naïve Bayesian classifier, logistic regression, and decision tree) to evaluate historical outpatient stroke medical data and signals from IoT sensors attached to the body to predict the probability of an impending stroke. Two optimization approaches, namely, the weighted sum rate maximization approach and the proportional fairness approach, are presented to achieve this goal. The proposed methods improved the average signal-to-noise ratio and proportional fairness [20].

Ge *et al.* demonstrated a system that collects data from sensors and used deep learning to evaluate and monitor patients' health data to predict disease and provide timely alerts in this work. Extensive analysis and experimental data are provided to show that the proposed strategy was safe and effective. When a patient uploads their health information, the system allows for precise access and confirmed deletion [21].

Lv *et al.* developed the theory of fuzzy function-based mean clustering algorithm using K-means and fuzzy theory in big data analysis technology. The results indicated that when the effective propagation probability was 100% and the value was between 0.01 and 0.05, it was close to the true result and has the least data delay. This study showed that using big data analysis techniques to enhance electric vehicle transportation

networks can significantly reduce network data transmission performance delay and adjust the route to effectively stop the spread of congestion [22].

Kaya *et al.* studied priority of speed and accuracy in health data and aimed to detect anomalies at the edge of IoT for effective management of big data. The data stream for age, gender, height, time, temperature, and weight is used in the analysis through applied Naïve Bayes, neural network, logistic regression, and random forest algorithms. The experimental results compared speed and accuracy and obtaining logistic regression (LR) algorithm provides great success in the IoT system. Machine learning (ML) algorithms are suitable for the IoT edge because they can make timely and efficient decisions in the health-care sector [23].

Ahmed *et al.* designed an approach based on neural networks, which aimed to diagnose and predict the epidemic. This model provided descriptive, diagnostic, predictive, and prescriptive analysis using big data analytics. This model was used to predict COVID-19 within the creation of a health monitoring platform. The results of the neural network-based model were also compared with the results of other machine learning methods. The neural network-based algorithm achieves high percentage accuracy without using any computationally expensive deep learning-based methodology [24].

Qin *et al.* used big data analytics to implement the multimedia-assisted student-centered learning model, which addresses critical policies to help educational institutions achieve this transformation in a more systematic way by clarifying teachers' instruction. When compared to alternative technologies, simulation results indicate that the proposed model improves student attention (97.3%), efficiency (90.1%), student retention rate (97.5%), engagement (98.2%), and learning outcomes (95.3%) [25].

3. DATA SET

The University, of Wisconsin, (USA) provided this breast cancer database. This dataset consists of 699, instance and 10, attributes, the target (class): (two for benign and four for malignant). The dataset contains (458) benign, class and (241) malignant class so that the data of dataset are imbalanced data. The description attributes with its values about this dataset can be summarized as follows:

Sample code number, clump thickness (1–10), uniformity, of cell size (1–10), uniformity of cell shape (1–10), marginal adhesion (1–10), single, epithelial cell size (1–10), bare nuclei (1–10), bland chromatin (1–10), normal chromatin (1–10), mitoses (1–10), and class (two for benign and four for malignant).

4. PROPOSED APPROACH

The dataset including information about the situation of patients, then will be processed using prediction approach such as Kernal Naive Bayes, Linear support vector machine (SVM), Coarse Tree, K-Nearest Neighbors, Cosine KNN, Coarse Gaussian SVMs and Fine Tree. Then, the performance of the algorithms will be evaluated using a suitable evaluation model such as k-fold cross-validation and percentage split approaches. The proposed approach includes many parts as shown in Fig. 1.

1. Breast cancer dataset: Including preparing and organizing the dataset to be accessible by the next step.

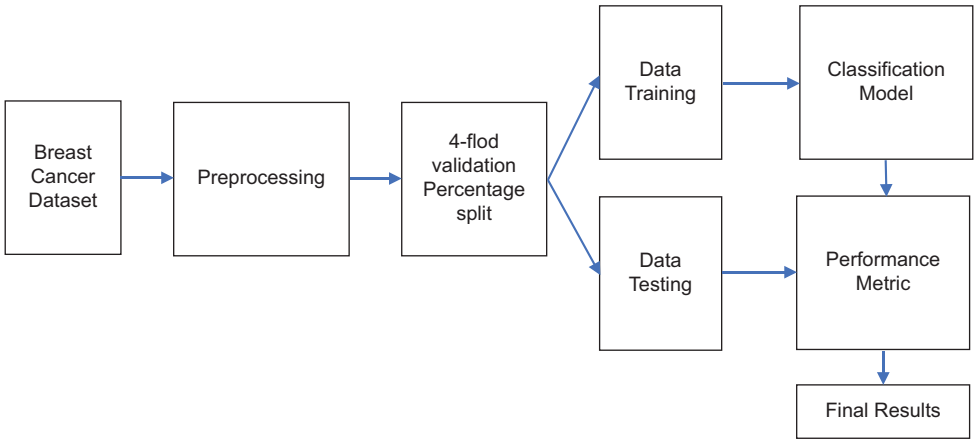

Fig. 1. Proposed approach for breast cancer dataset.

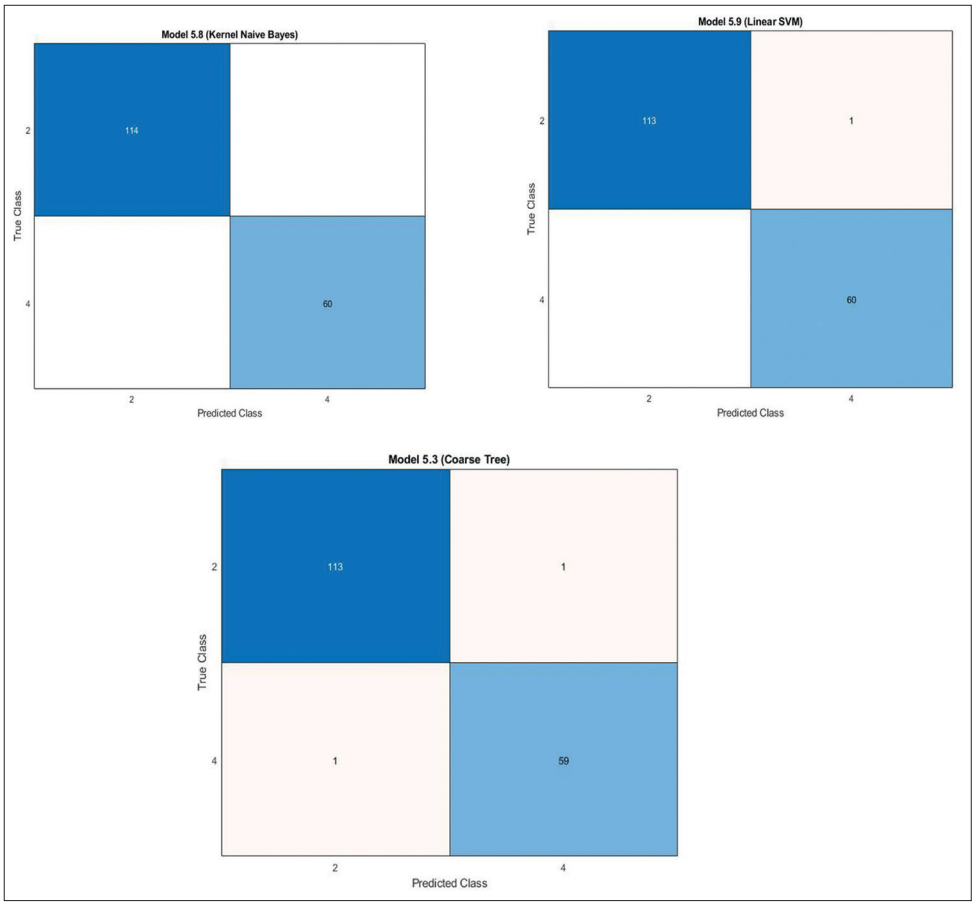

Fig. 2. Confusion matrix of breast cancer (percentage split).

2. Data preprocessing: include three parts: outlier detection, missing values and normalization which can perform by using many functions provided with MATLAB classification learner.
3. Feature selection: It is the process of choosing the essential variables to improve accuracy. The important and affected features that are required to increase accuracy are chosen in this work.
4. Data splitting: Breast cancer dataset was divided into: training and test sets using 4-flod validation and percentage split. This splits randomly with 25% held out for testing and 75% for the training.
5. Machine learning models: Coarse DT has few leaves and distinguishes between classes with coarse distinctions and a maximum of four splits, Linear SVM, Kernel Naïve Bayes for percentage split and Fine Tree, Cosine KNN, Coarse Gaussian SVM and Subspace KNN for 4-fold cross validation in breast cancer dataset.

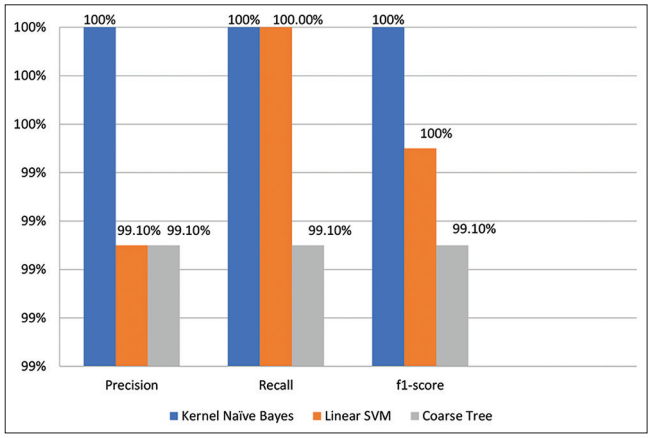

Fig. 3. Comparison of kernel NB, linear support vector machine, coarse tree algorithms.

5. DISCUSSION AND ANALYSIS

The of breast cancer confusion matrix is implemented in the case of percentage split which is used to calculate different

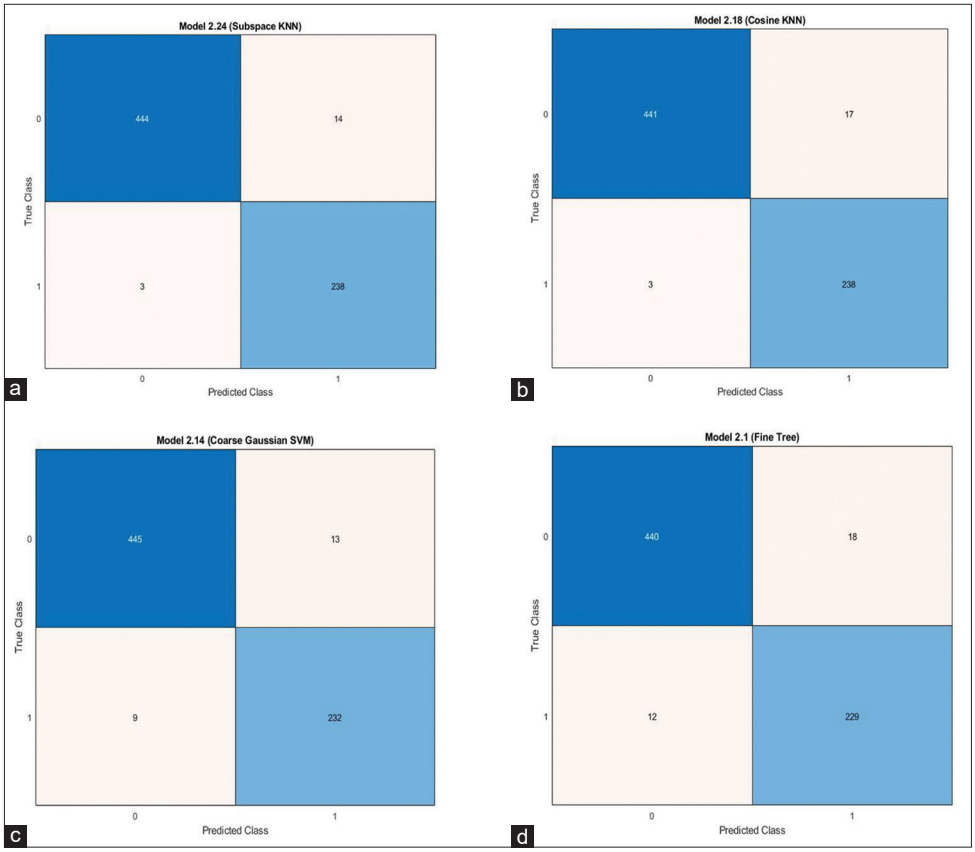

Fig. 4. Confusion matrix of breast cancer (4-fold cross validation).

performance metrics such as precision, Recall and f1-score as shown in Fig. 2.

The precision, recall and f1-score are measured for Kernel Naïve Bayes, Linear SVM, coarse tree classifiers respectively. Fig. 3 shows the performance comparison values of Kernel Naïve Bayes, Linear SVM, and Coarse Tree classifier algorithms in case of percentage split.

In this work, different classifier algorithms are applied on breast cancer dataset. The Kernel Naïve Bayes is adapted to classify the breast cancer and the outcomes of the experimented results are compared the precision, recall and f1-score of many classifier algorithms like Coarse Tree and Linear SVM. So that the obtained results are 100% for all precision, recall and f1- score for Kernel Naïve Bayes approach.

Applying different classification algorithms on breast cancer dataset leading that the subspace KNN is adapted to classify the breast cancer in case of 4-fold cross validation. Fig. 4 shows the confusion matrix which used to compute the different performance metrics such as accuracy, sensitivity, specificity and f-measure.

Precision, recall, and f1-score are measured for Subspace KNN, Cosine KNN, Coarse Gaussian SVM and Fine Tree classifiers. Fig. 5 shows the performance comparison values of Subspace KNN, Cosine KNN, Coarse Gaussian SVM and Fine Tree classifier algorithms in case of 4-fold cross validation. This figure indicated a better performance in case of subspace KNN compared to other algorithms.

Table 1 shows the outcomes of the proposed approach compared with other studies that used this dataset. The classification algorithms used in this study show that the obtained performance is 100% for precision, recall, and f1-score compared to reference [7]. In addition, the performance is 97.2%, 99.1%, and 98.1% for precision, recall, and f1-score compared to reference [8].

TABLE 1: Results are compared to previous results in breast cancer dataset

This study	Other study	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-scaore (%)
Applying percentage split evaluation technique	Dhanya *et al*. [7]	Used ensemble techniques	97.86	-	-	-
	Classification via applied approach		-	100	100	100
Applying 4-fold cross validation	Bayrak *et al*. [8]	Used support vector machine technique		96	95.7	-
	Classification via applied approach			97.2	99.1	98.1

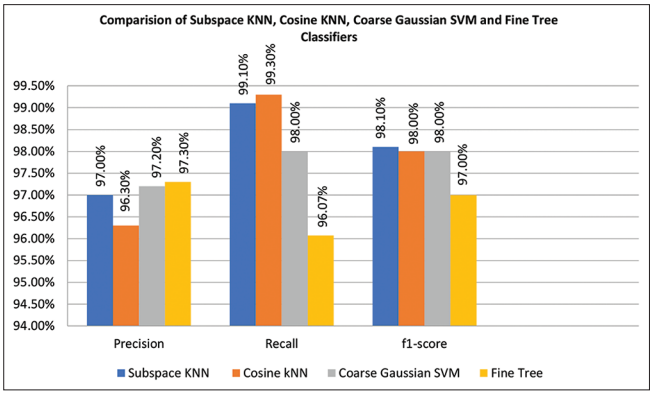

Fig. 5. Metrics of breast cancer (4-fold cross validation).

6. CONCLUSIONS

Efficient big data analytical approaches and machine learning algorithms are utilized for accurate prediction and decision making. So that different machine learning algorithms are being implemented for predicting breast cancer. The machine learning techniques are using comparisons based on important performance metrics such as precision, recall, and f1-score. Machine learning (disease prediction) is a suitable technique that assists in the early detection of disease, may aid practitioners in diagnosis decision-making, and also gives accurate predictions. Many machine learning algorithms applied on Wisconsin breast cancer (WBC) dataset that generate best performance accuracy for diagnosis and prediction the WBC dataset, in which the Kernel Naïve Bayes gives 100% as precision in percentage split method, the Coarse Gaussian SVM is adapted to classify the breast cancer in case of 4-fold cross validation which gives 97.2% as precision.

REFERENCES

[1] P. Raj, T. Poongodi, B. Balusamy and N. Khari. "*The Internet of Things and Big Data Analytics Integrated Platform and Industry Use Cases*".1st ed. Routledge, United Kingdom, 2020.

[2] L. Wang and R. Jones. "Big data analytics in cyber security: Network traffic and attacks*". Journal of Computer Information Systems*, vol. 61, pp. 1-8, 2020.

[3] Z. A. Dagdeviren and O. Dagdeviren. "BICOT: Big data aanlysis approaches for clustering cloud based IOT systems". *European Journal of Science and Technology*, vol. Special Issue 26, pp. 395-400, 2021.

[4] K. Ding and P. Jiang. "RFID-based production data analysis in an IoT-enabled smart Job-shop". *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no.1, pp. 128-138, 2018.

[5] S. Sawalha and G. Al-Naymat. "Towards an efficient big data management schema for IoT". *Journal of King Saud University-Computer and Information Science*, vol. 34, pp. 7803-7818, 2021.

[6] N. Arhab, M. Oussalah and M. S. Jahan. "Social media analysis of car parking behavior using similarity based clustering". *Journal of Big Data*, vol. 9, p. 74, 2022.

[7] R. Sharma, D. K. Sharma, D. Bhatt and B. Thai Pham. "*Big Data Analysis for Green Computing Concepts and Applications*". 1st ed. Routledge, United Kingdom, 2022.

[8] H. Nasiri, S. Nasehi and M. Goudarzi. "Evaluation of distribute stream processing framework for IOT applications in smart cities". *Journal of Big Data*, vol. 6, p. 52, 2019.

[9] F. Ullah, H. Naeem, S. Jabbar, S. Khalid, M. A. Latif, F. Al-Turjman and L. Mostarda. "Cyber security threats detection in internet of things using deep learning approach". *IEEE Access*, vol. 7, pp. 124379-124389, 2019.

[10] A. K. Sangaiah, A. Thangavaelu and V. M. Sundaram. "*Cognitive Computing for Big Data Systems Over IoT Frameworks, Tools and Applications*. Springer, Berlin, 2018.

[11] X. Nie, T. Fan, B. Wang, Z. Li, A. Shankar and A. Manickam. "Big data analytics and IoT in operation safety management in under water management". *Computer Communications*, vol. 154, pp. 188-196, 2020.

[12] A. Akbar, G. Kousiouris, H. Pervaiz, J. Sancho, P. Ta-Shma and F. C. K. Moessner. "Real-time probabilistic data fusion for large-scale IoT application". *IEEE Access,* vol. 6, pp. 10015-10027, 2018.

[13] D. Serpanos and M. Wolf. "*Internet-of-Things (IoT) Systems Architectures, Algorithms, Methodologies*". Springer, Berlin, 2018.

[14] S. A. Shah, D. Z. Seker, S. Hameed and D. Draheim. "The rising role of big data analytics and IoT in disaster management: Recent advances, taxonomy and prospects". *IEEE Access*, vol. 7, p. 54595-54614, 2019.

[15] M. Babar, M. D. Alshehri, M. U. Tariq, F. Ullah, A. Khan, M. Irfan Uddin and A. S. Almasoud. "Energy-harvesting based on internet of things and big data analytics for smart health monitoring". *Sustainable Computing: Informatics and Systems*, vol. 20, pp. 155-164, 2018.

[16] S. Singh and A. Yassine. "*IoT big data analytics with fog computing for household energy management in smart grids". Smart Grid and Internet of Things*. Springer, Berlin, pp. 13-22, 2018.

[17] S. A. Shah, D. Z. Seker, M. M. Rathore, S. Hameed, S. B. Yahia

and D. Draheim. “Towards disaster resilient smart cities: Can internet of things and big data analytics be the game changers? *IEEE Access*, vol. 2, p. 1-19, 2019.

[18] L. Syed, S. Jabeen, S. Manimala and A. Alsaeedi. “Smart health framework for ambiet assisted living using IoMT and big data analytics techniques”. *Future Generation Computer System*, vol. 101, pp. 136-151, 2019.

[19] C. Li. “Information processing in Internet of Things using big data analytics”. *Computer Communications*, vol. 160, pp. 718-729, 2020.

[20] M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi and J. M. Elmirghani. “Patient-centric HetNets powered by machine learning and big data analytics for 6G networks”. *IEEE Access*, vol. 8, pp. 85639-85655, 2020.

[21] C. Ge, C. Yin, Z. Liu, L. Fang, J. Zhu and H. Ling. “A privacy preserve big data analysis system for wearable wireless sensor network”. *Computers and Security*, vol. 96, p. 101887, 2020.

[22] Z. Lv, L. Qiao, K. Cai and Q. Wang. “Big data analysis technology for electric vehicle networks in smart cities”. *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, pp. 1807-1816, J 2020.

[23] S. M. Kaya, A. Erdem and A. Güneş. “A smart data pre-processing approach to effective management of big health data in IoT edge”. *Smart Homecare Technology and TeleHealth*, vol. 8, pp. 9-21, 2021.

[24] I. Ahmed, M. Ahmad, G. Jeon and F. Piccialli. “A framework for pandemic prediction using big data analysis”. *Big Data Research*, vol. 25, p. 100190, 2021.

[25] T. Qin, P. Poovendran and S. BalaMurugan. “Student-centered learning environments based on multimedia bi data analytics”. *Arabian Journal for Science and Engineering*, vol. 48, p. 1-11, 2021.