International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 5, October 2021, pp. 4439~4448
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i5.pp4439-4448

Security aware information classification in health care big data

Snehalata K. Funde, Gandharba Swain
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India

Article Info

Article history:

Received Jun 15, 2020
Revised Apr 8, 2021
Accepted Apr 27, 2021

Keywords:

Big data
Classification
Healthcare
Sensitive data
WBS

ABSTRACT

These days e-medical services frameworks are getting famous for taking care of patients from far-off spots, so a lot of medical services information like the patient's name, area, contact number, states of being are gathered distantly to treat the patients. A lot of information gathered from the different assets is named big data. The enormous sensitive information about the patient contains delicate data like systolic BP, pulse, temperature, the current state of being, and contact number of patients that should be recognized and sorted appropriately to shield it from abuse. This article presents a weight-based similarity (WBS) strategy to characterize the enormous information of health care data into two classifications like sensitive information and normal information. In the proposed method, the training dataset is utilized to sort information and it comprises of three fundamental advances like information extraction, mapping of information with the assistance of the training dataset, evaluation of the weight of input data with the threshold value to classify the data. The proposed strategy produces better outcomes with various assessment boundaries like precision, recall, F1 score, and accuracy value 92% to categorize the big data. Weka tool is utilized for examination among WBS and different existing order procedures.

Corresponding Author:

Snehalata K. Funde
Department of Computer Science and Computer Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram-522502, Guntur, Andhra Pradesh, India
Email: snehalatafunde@gmail.com

1. INTRODUCTION

Nowadays data in various fields like government organizations, health care systems, military, and banking sectors are growing exponentially. As the data is huge in amount, it needs to be stored in digital form. In earlier years, even though the size of data does not matter, still the inflow from where the information comes from and the structure of that information was restricted. In today's world, the situation has changed and a big amount of data can be fetched from enormous sources and in a variety of its formats. The various tools like hadoop distributed file system (HDFS) and MapReduce are used to store such a big amount of data [1], [2]. The role of big data in the NoSQL database is in an enormous form to achieve high performance and accuracy over conventional databases. It is challenging to handle such a big amount of data using rows and columns format when the input data is in an unorganized format. NoSQL databases effectively handle such kinds of unstructured data [3]. One of the well-known sources of big data is logs generated through various web and desktop applications. The data coming from these sources are having types like well-organized data, unorganized data, or unstructured data. Big data having different phases, which are termed lifecycle phases of big data [4]. The first phase is the Collection of the input data from legal and authorized sources and gets it available as an input for the next phase of the lifecycle. The second phase is to store the collected data using trusted functions. This is the phase where the chances to get big data

access by the attacker are very high. Big data analysis and processing is the next step in the lifecycle phases. In this phase, we need to maintain the integrity of the data, which is being processed. The last phase is to generate meaningful knowledge from the analyzed data. This meaningful information is helpful to make the decisions for the business models of the organizations. The generated knowledge is used as sensitive data and various security tools and techniques can be used to secure this generated sensitive data.

In reality, the problem of analyzing big data is there for many years as data creation is simpler than finding beneficial information from it. Even if there is much fast processing equipment, processors are available in the market, huge data processing is still a tedious task [5]. Since the conventional techniques are unable to handle the processing load of the big data, so algorithms are created in such a way that they should be compatible with MapReduce as well as SailFish frameworks to handle big data [6]. There are some areas like transport systems where the knowledge produced by processing the big data is used. To increase the performance of this kind of system organizations and government needs processed data for decision-making. The accuracy of the decision-making increases proportionally with the number of training datasets as input [7]. Modernizing the hardware of the system is necessary along with various platforms to analyze and extract meaningful knowledge from the big data. Optimal matching for hardware and software within a specific time to analyze big data is one of the challenges. Selecting the right platform to increase the scalability and performance of the system, which leads to accurate results generation. The generated results are also termed sensitive data. This sensitive data can be used for further evaluation as an input to other systems [8].

Big data require new mining strategies for processing the data. The definition of 3Vs of big data implies that the data are huge in amount, the data might be created in hurry, and the information will exist in multiple types as captured from extraordinary resources [9], [10]. One of the trusted sources for input data in the modern era is the use of intelligent objects using the internet of things (IoT). These devices have collaborated with internet and the acts as a mediator between users and other gadgets or appliances. The amount of data generated during communication can be used as big data [11]. The healthcare organization has generated a big quantity of information from the report keeping, compliance, and patient associated statistics [12].

The cryptographic system requires multiple keys for the encryption process so instead of encryption technique if the classification approach is used to categorize data into normal attributes and sensitive attributes category then encryption techniques can be applicable on only sensitive data [13]. Big data are analyzed not only on-premises/servers but also on cloud platforms. The data are stored and can be analyzed on the cloud where frameworks are provided to do so [14]. In the digital world, data need to be digitized to enhance the healthcare system by minimizing the charges and analyzing many records within a short period successfully [15]. It needs an advanced approach to deal with huge data. The main thing in health care usually consists of sensitive facts, which want to be covered from any unanticipated operations or record retrievals [16], [17]. Moreover, analyzing personal information or dataset is problematic because many related facts may additionally comprise sensitive things [18], [19].

Health care data costs more as it carries sensitive data of the patient. According to the Ponemon Institute report, the average healthcare record breach price is almost $380. In many cases, data may be lost due to a lack of secured data structures, machines, and data storage strategies. To permit employees to have data records on their devices is never so easy. Privacy of data is considered a primary issue in big data analysis [20], [21]. The healthcare domain is widely used as one of the trusted sources of datasets. The biosensor device can be used to collect health information from the patient. The sensor is applicable using the external part on the body and the report generated will be stored in the available databases. The results generated by the sensor are the mixed format input data. We can classify those input data into normal and sensitive data. To classify big data, proper classification techniques are required which will give accurate results in less amount of time.

There are many classification techniques introduced by different people in recent days. Some of the techniques are as below:

- Naive bayes classifier: The naive bayes classifier is used to classify the input data from the health care dataset. The information related to blood pressure, and pulse rate. The limitation of the Naive based classifier is if the dataset size is small, the precision will decrease [22].
- Improved support vector machine classifier (SVM): Improved SVM is used to classify big data by adding extra functionalities like a weighted Euclidean distance, radial integral kernel functioning existing SVM. It is suitable for multi duplicated samples and a large amount of data [23].
- Bagging-based naive bayes trees: This classification technique is used to classify big data by combining two approaches; i) bagging ensemble, and ii) naïve bayes, and the decision tree approach. It can classify big data in less amount of time [24].

WBS classification technique, particularly designed for those applications concerning the information security of big data as an input. The investigation of the proposed algorithm is essential, as an

accurate classification of datasets into two categories sensitive and non-sensitive is needed in the view of the security of data in various domains.

2. RELATED WORK

Pham and Prakash [24] proposed the bagging-based naive bayes trees (BAGNBT) approach. It is used for landslide vulnerability categorization in Viet Nam. This method was approved utilizing the tests such as the Chi-square test and statistical indexes. For the correlation, different analysis models were chosen in this research. The output shows that the novel model with the area under the receiver operating the characteristic curve (AUC) [25] (0.834) resulted better with a comparison of rotation forest-based naïve bays trees (RFNBT) (0.830). This demonstrates that the technique is a promising and better elective technique for landslide vulnerability assessment. Techniques like the AUC, statistical pointers, along the Chi-Square test have been used to validate the models in the current test. AUC is used as a standard for validating prototypes. "Sensitivity" and B100-specificity" [26] parameters are used for plotting the AUC. For value 1, models are considered as immaculate. Models are labeled as good with greater AUC value. For the approval of the model different performance measures like root mean squared error (RMSE), Kappa (κ), and Accuracy (ACC) are used [27]. Landslide causing parameters, in particular incline, separation to deficiencies, bend, street thickness, profile bend, perspective, plan bend, waterway thickness, lithology, height, separation to streets, separation to rivers, precipitation, flaw thickness, and area usage, were chosen for the evaluation of landslide vulnerability. For examination purposes, maps of these components were created. BAGNBT approach includes four main steps as given below:

a. Dataset generation: Mainly two necessary datasets were created with this step. First is the training dataset and the second is validating the dataset. For the training dataset, they have used 70% of the datasets with rockfall and non-rockfall values while for validating the datasets they have used 30% remaining data. Boolean values 1 and 0 were used for rockfall and non-rockfall simultaneously.
b. Novel model training: The training dataset was used for training the hybrid model in this approach. In this phase, to optimize the provided dataset for the categorization BAG ensemble was used. 11 iterations were identified to match the high accuracy of the model named bagnbt. along with it, a classifier like Naïve Bays Tree (NBT) was used to label the non-landslide and landslide classes for the geographical forecast of the landslide using upgraded training data items. In the last step, a combo of BAG was recycled for consolidation produced classifiers of naïve bays tree to develop a brand new model.
c. Model validation: The novel model was approved utilizing different strategies such as AUC, and the Chi-Square test.
d. Developing landslide weakness maps: Landslide map vulnerability was created utilizing the NBT, SVM, RFNBT, and BAGNBT models.

Rockfall vulnerability evaluation was performed at most sensitive cities where landslides happen very frequently like Viet Nam using the suggested method with the approach of a combination of the Naive Based Trees as a classifier with the BAG collaborative. The statistical pointers, AUC, and chi-square methodology has functioned for testing. Estimation and relationship consequences of the proposed system prove that the BAGNBT model has a prominent demonstration for rockfall vulnerability analysis of AUC whose value is equal to 0.834 in examination with the RFNBT value which is equal to 0.830. It classifies data moderately. As it contains multiple iterations for classifying data, its weakness is a bit time-consuming.

Lakshmanaprabu *et al.* [28] introduced a method for the classification of data with a random forest (RF) classifier. The RF classifier is used for the analysis of particular big data collected from various sources in this paper. Figure 1 shows the data collection process from various data sources. Particularly they used patient data including many health problem parameters. For the proper classification of data taken from the healthcare, database author has utilized the improved dragonfly algorithm [29]. With the assistance of ideal features, a selected RFC classifier [30] is applied to characterize the healthcare data. Output values got from the result of precision are at most 94.2 as per execution. So distinctive measures are used and compared with existing strategies to check the viability of the techniques. Author has used actual health care data as test data and web information as training data. This work has different phases such as information accumulation, observing, splitting, making digital, controlling, and support. After data accumulation, the featured those are highlighted are taken out with the help of the dragonfly algorithm and given to RF classifier for further classification of data according to the user's requirement.

The RF classifier technique performs well for classifying the data as it does not have much effect on disorders of big data, variations in data with huge amounts. It is based on a tree generation approach in which many trees are generated for the estimation of the result to be chosen as a better one. RF develops a random sample of the data and identifies the key organization of features for making a choice tree. Figure 2 demonstrates the RF structure. RF makes a case of the data and checks with many decision trees generated before for the finalization of optimal solutions.

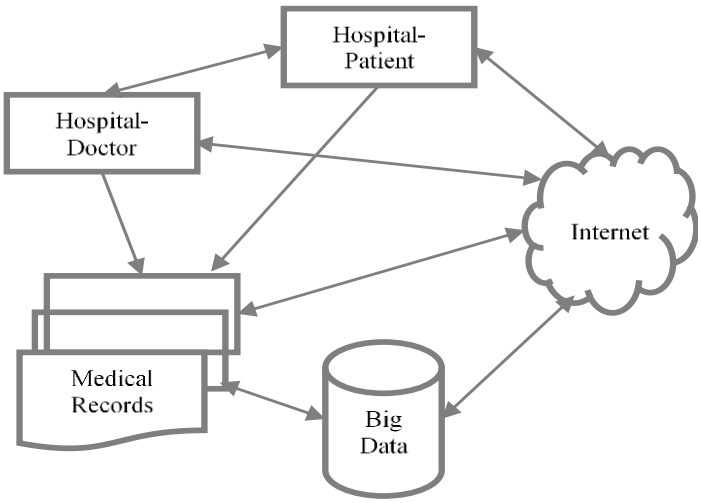

Figure 1. Healthcare data

The training dataset is provided to substitute bootstrap tests for generating every choice tree. Amid the enhancement of a decision tree at each node division, an irregular subset of few components is browsed the first-factor set and the best division in light of these few factors is used.

RF classifier mainly referred to three key parameters for the classification of big data as below:

– Node size is taken that is not similar to the comparison of decision trees.
– Trees count up to 500 trees is commonly a proper choice.
– The count of the number of sampled predictors to be tried at each split would have all the earmarks of being a key parameter that should affect how well RF performs. For error cases, the following (1) is used.

$$RandomForestError = RandomForest\, r_1, r_2\, (F(v_1, v_2) < 0) \quad (1)$$

where, r1, r2 are random values and v1 and v2 are vectors.

Quality and relationship parameters measure the accuracy of each classifier and dependency between classifiers. Different characteristics used at the phase of the base classifier return the quantity of data that is arbitrarily picked from the initial organization of the properties particularly at each node of the base decision tree. The most extraordinary probability strategy is used to secure perfect results in each tree so that it can be helpful in RF classification. In this research, accuracy gets over 90% and 95%. The weakness of the proposed calculation is computationally moderate as the result of the huge database.

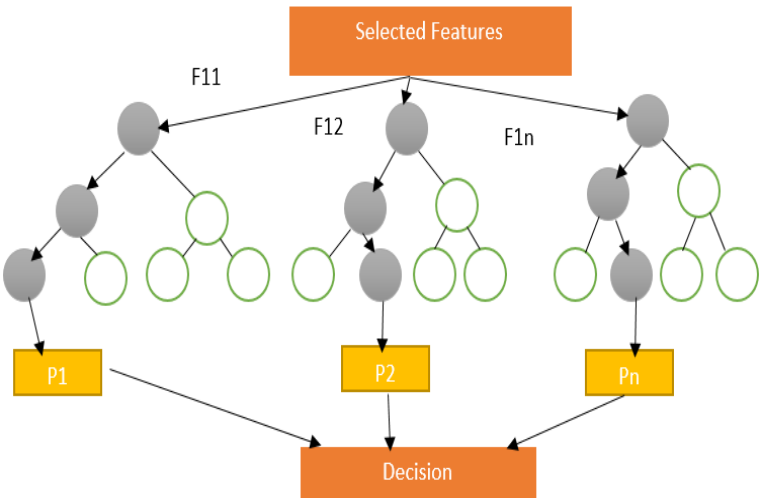

Figure 2. Sample structure of RF

In the existing systems, the classification techniques like SVM, RF, BAGNBT are used for data classification, and the results are analyzed in terms of precision, recall, F1 score, and accuracy parameters. In the proposed system, weight is calculated, and it is compared with the threshold value to decide on normal data and sensitive data. The generated results for the proposed algorithm and existing systems are compared based on kappa parameters, which shows superior results for the proposed system.

3. PROPOSED METHOD

For the proposed algorithm multiple attributes from network packet instances of health care data system are considered as input (step 1). For the classification of data into two categories, sensitive data and normal data background knowledge is used (step 2). The proposed technique is based on an instance-based method so it fetches attributes per instance generated from the network. After that background knowledge is applied to each attribute and tokenization is done for every attribute. With the help of background knowledge each input data mapping is performed (step 3). Weight is calculated for each mapping and compared with Threshold (step 4). If the resultant weight value is greater than or equal to the threshold, then that attribute will be considered as a sensitive attribute otherwise it will be considered as a normal attribute (step 5). Finally, in the end, two lists will be created for sensitive and normal data.

3.1. Algorithm

The Proposed WBS technique collects data from the health care system with the help of different types of sensors attached to the patient body to measure different health conditions of the patient. Furthermore, to categorize data, the patient database is taken as an input to the system, which contains multiple records. Each record has multiple patient attributes such as patient name, address, mobile number, Blood pressure. In the algorithm, R[i] shows patient records and A[j] shows attributes of every patient record. In this system, T shows a training dataset with the E number of examples to compare patient data for sensitive and non-sensitive attributes. As shown in (2) is used for the tokenizing record into multiple attributes and (3) is used for calculating the weight of the attribute with the similarity distance method. This outcome will be in the form of a set of sensitive attributes SA[j] and a set of non-sensitive attributes NSA[j].

Input: R[i], n number of records where i=1…n.
A[j], n number of attributes from each record, where j=1...n.
T, Training set with ||E|| examples.

Output: SA[j] a set of sensitive attributes.
NSA[j] set of non-sensitive attributes.

Step 1: Tokenize each record into attributes.

$$\text{Tokenize}(\text{R[i]}) = \text{A[1]}, \text{A[2]} \ldots \text{A[n]} \quad (2)$$

This step splits Patient records into separate attributes.

Step 2: Calculate the weight of each attribute by checking the similarity of the attribute with given training set examples.

$$\text{weight}(\text{A[j]}) = \text{similarityDistance}(\text{A[j]}, \text{T}) \quad (3)$$

Similarity distance is calculated by using the cosine similarity measure for finding the similarity between two objects or vectors. It is calculated using formula $\text{Cos}\,\theta = \frac{\vec{A}\cdot\vec{B}}{||\vec{A}||\cdot||\vec{B}||}$.

Step 3: The System validates attribute weight with a defined threshold (Th) value. (Here, used Th=0.01)

Step 4**:** if(weight $\geq$ Th)

Append the attribute to sensitive list SA[j]

else

Append the attribute to non-sensitive list NSA[j]

Step 5: Return {SA[j], NSA[j]}

Based on the above statement, the algorithm classifies each attribute either it is sensitive attributes or non-sensitive attributes.

3.2. Flow diagram

Figure 3 shows the flow of the WBS technique for classifying data into two categories sensitive data and non-sensitive data.

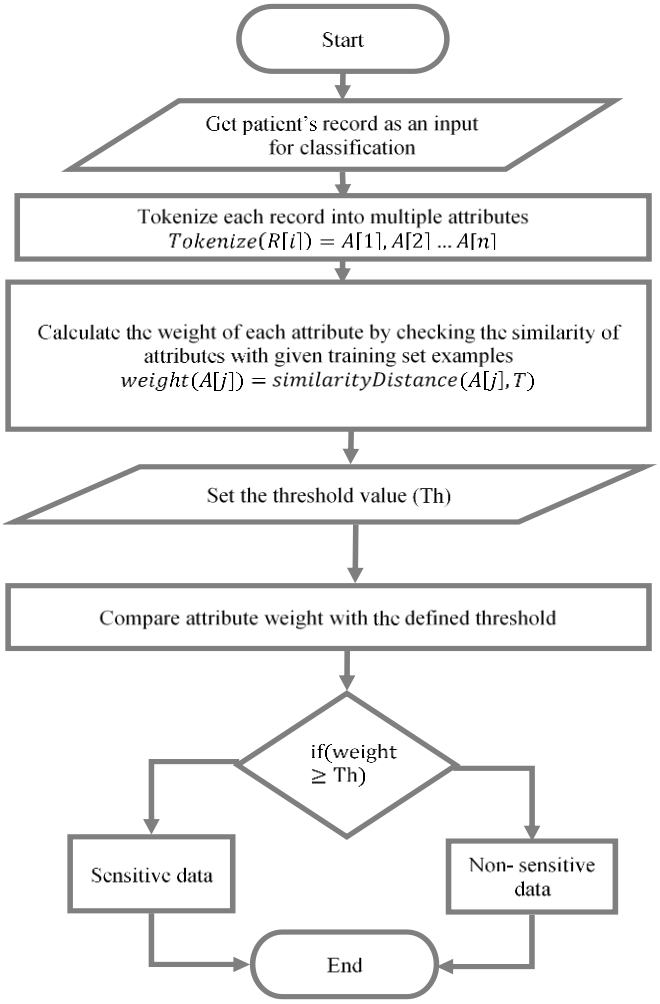

Figure 3. The flow of weight-based similarity technique

4. RESULTS AND DISCUSSION

This proposed work has been implemented by using a 64-bit Windows 10 operating system, NetBeans as an Integrated Development environment, Weka tool to compare kappa parameters, and Java technology. The hardware used for this work is an Intel i7 processor with 8 GB of RAM. For generating results, we have considered one lac instance of records, which are collected from real-time data generated by different sensor devices by the hospital. The data generated used for training and result generation purpose and the values of the attributes are real-time for systolic BP, diastolic BP, heart rate, total cholesterol, Cholesterol_LDL, Cholesterol_HDL, stress, Random Sugar, QT_Interval, PR Interval, oxy_saturation, and HB. Part of the sample dataset of the patient for classification considered as input is displayed in below Figure 4. Figure 5 shows classified data into two categories normal data and sensitive data.

The proposed technique is compared with existing classification techniques such as support vector machine (SVM), random forest (RF), Bagging-based Naive Bayes Trees (BAGNBT. The comparison parameters are; i) Precision, ii) Recall, iii) F1 Score, and iv) Accuracy. These parameters are defined in (4)-(7), respectively.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive}+\text{False Positive})} \quad (4)$$

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive}+\text{False Negative})} \quad (5)$$

$$\text{F1 Score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision}+\text{Recall})} \quad (6)$$

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{(\text{True Positive} + \text{False Negative} + \text{True Negative} + \text{False Positive})} \quad (7)$$

Here, comparison parameters are defined based on positive and negative class strategy. Positive class refers to the set of sensitive data items from patient record. Negative class refers to a set of normal data items from patient record. True positive is the outcome where the classification algorithm accurately predicts a positive class. True negative is the outcome value where the algorithm predicts the negative class accurately. False positive is the outcome where the classification approach wrongly predicts positive class. False negative is the outcome where the classification technique incorrectly predicts the negative class. Figure 6 is a graphical representation of precision for existing techniques and proposed algorithm (WBS) where WBS has higher Precision (0.92) as compared to the existing techniques.

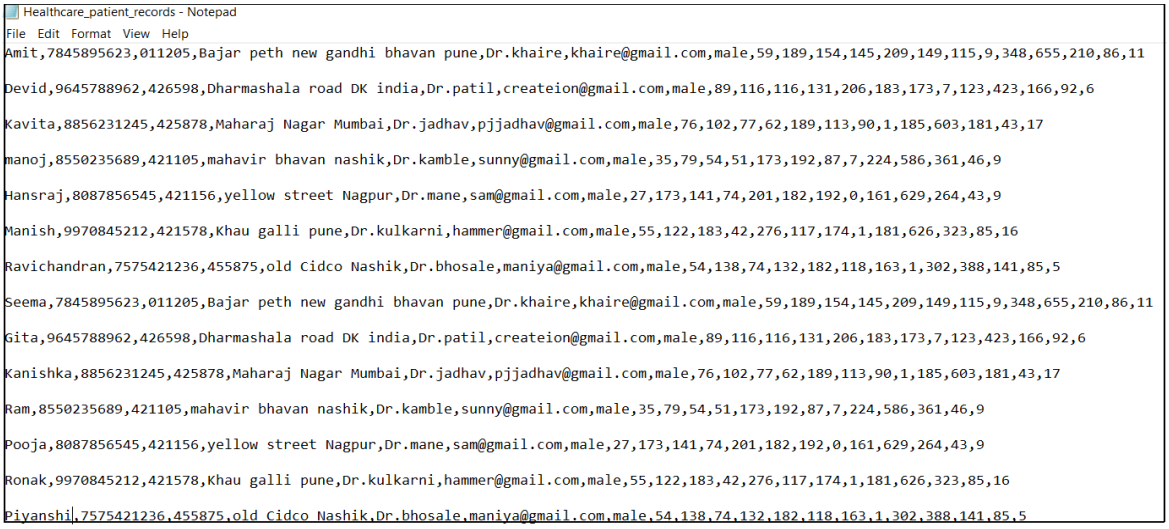

Figure 4. Sample input dataset

Output (Classified dataset):

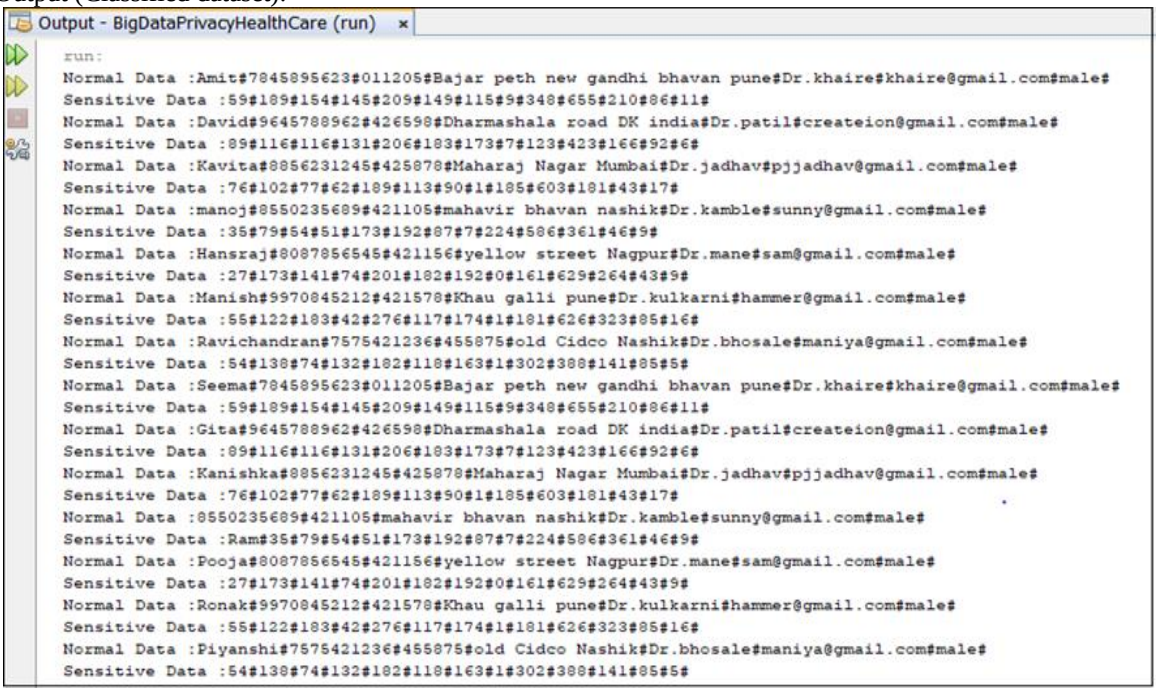

Figure 5. Classified sensitive data and normal data

Figure 7 is a graphical representation of Recall for existing techniques and proposed algorithm (WBS) where WBS has a higher value for Recall (0.91) as compared to the existing techniques. Figure 8 is a graphical representation of the F1 score for existing techniques and proposed algorithm (WBS) where WBS has a higher F1 score (0.89) as compared to the existing techniques. Figure 9 is a graphical representation of Accuracy for existing techniques and proposed algorithm (WBS) where WBS has higher Accuracy (92%) as compared to the existing techniques.

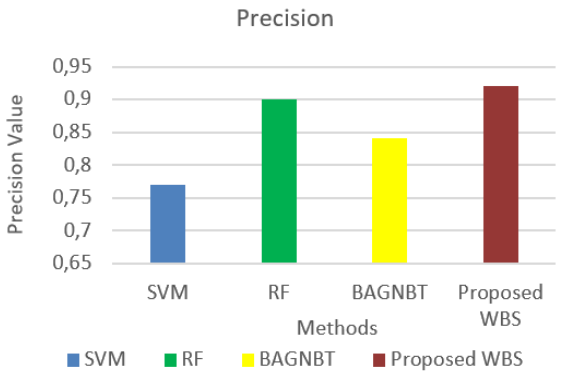

Figure 6. Precision comparison with existing algorithms

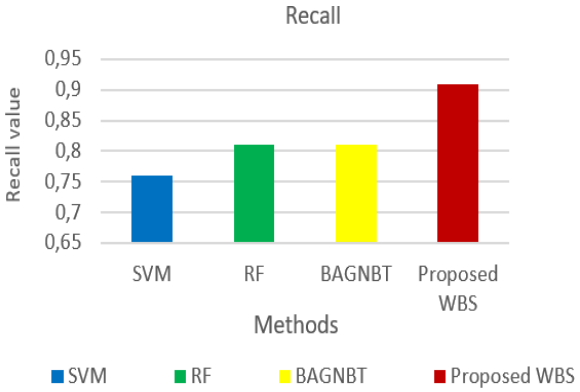

Figure 7. Recall comparison with existing algorithms

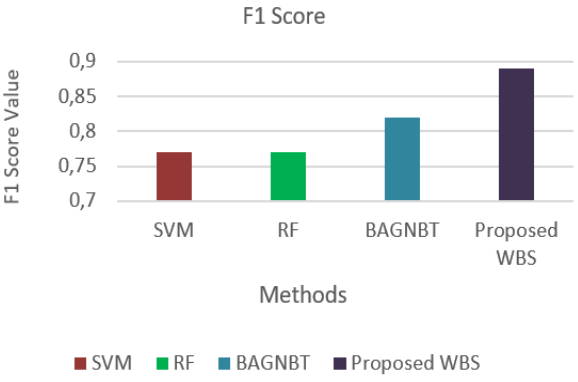

Figure 8. F1 Score comparison with existing algorithms

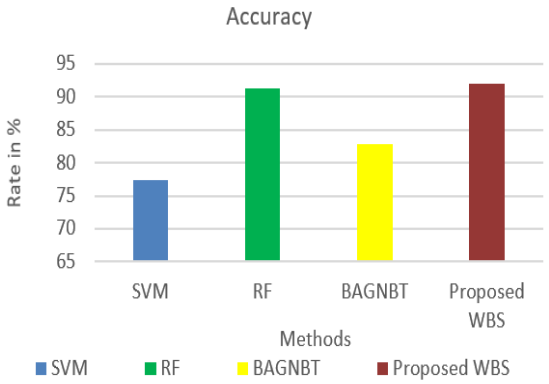

Figure 9. Accuracy comparison with existing algorithms

5. CONCLUSION

This article proposed a technique based on weight-based calculation for the classification of health care big data. It has three main steps data extractions, mapping of input data with background knowledge, and evaluation of the weight of input data with a threshold to classify data. In this technique, the health care dataset is properly classified using background knowledge provided for the data. Experimental results are estimated by using the WEKA tool. The proposed technique is compared with many existing techniques using different comparison parameters like Precision, Recall, F1 score, and accuracy. It performs better with an accuracy of 92% for classifying sensitive data as compared to existing techniques NB (82.76%), RF (91.20%), SVM (77.30%). As the proposed algorithm provides more accuracy to medical data, health care applications can categorize and store sensitive data more securely.

REFERENCES

[1] Y. Gao, X. Chen, and X. Du, "A Big Data Provenance Model for Data Security Supervision Based on PROV-DM Model," *IEEE Access*, vol. 8, pp. 38742-38752, 2020, doi: 10.1109/ACCESS.2020.2975820.

[2] R. Chintala, M. R. Narasinga Rao, and S. Venkateswarlu, "Review on the Security Issues in Human Sensor Networks for Healthcare Applications," *International Journal of Engineering and Technology*, vol. 7, no. 2.32, 269-274, 2018, doi: 10.14419/ijet.v7i2.32.15582
[3] R. S. Raghav, J. Amudhavel, and P. Dhavachelvan, "A Survey of Nosql Database for Analyzing Large Volume of Data in Big Data Platform," *International Journal of Engineering and Technology (IJET)*, vol. 7, no. 32, pp. 181-186, 2018, doi: 10.14419/ijet.v7i2.32.15563
[4] B. S. Alladi, and S. Prasad, "Big Data Life Cycle: Security Issues, Challenges, Threat and Security Model," *International Journal of Engineering and Technology*, vol. 7, no. 3, pp. 100-103, 2018, doi: 10.14419/ijet.v7i1.3.9666.
[5] B. Yang, and T. Zhang, "A Scalable Meta-Model for Big Data Security Analyses," *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, New York, NY, USA, 2016, pp. 55-60, doi: 10.1109/BigDataSecurity-HPSC-IDS.2016.71.
[6] D. Radhika, and D. Aruna Kumari, "Adding Big Value to Big Businesses: A Present State of the Art of Big Data, Frameworks and Algorithms," *ICT Based Innovations*, vol. 653, pp. 171-184, 2017.
[7] N. V. Ramana, P. V. M. Seravana Kumar, and P. Nagesh, "Analytic Architecture to Overcome Real Time Traffic Control as an Intelligent Transportation System using Big Data," *International Journal of Engineering and Technology*, vol. 7, no. 18, pp. 7-11, 2018, doi: 10.14419/ijet.v7i2.18.10772.
[8] A. Londhe, and P. P. Rao, "Platforms for Big Data Analytics: Trend towards Hybrid Era," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, 2017, pp. 3235-3238, doi: 10.1109/ICECDS.2017.8390056.
[9] R. Lu, H. Zhu, and X. Liu, "Towards Efficient and Privacy Preservation Computing in Big Data Era," *IEEE Network*, vol. 28, no. 4, pp. 46-50, 2014, doi: 10.1109/MNET.2014.6863131.
[10] D. Ardagna, C. Cappiello, W. Sama, and M. Vitali, "Context-Aware Data Quality Assessment for Big Data," *Future Generation Computer Systems*, vol. 89, pp. 548-562, 2018, doi: 10.1016/j.future.2018.07.014.
[11] N. K. Rao, and G. Swain, "A Systematic Study of Security Challenges and Infrastructures for Internet of Things," *International Journal of Engineering and Technology*, vol. 7, no. 4.36, pp. 700-706, 2018, doi: 10.14419/ijet.v7i4.36.24226
[12] S. Nazir, S. Khan, H. Khan, S. Ali, I. Magarino, R. Atan *et al.*, "A Comprehensive Analysis of Healthcare Big Data Management, Analytics and Scientific Programming," *IEEE Access*, vol. 8, pp. 95714-95733, 2020, doi: 10.1109/ACCESS.2020.2995572.
[13] S. Funde, and G. Swain, "Big Data Storage and Access: A Security Analysis," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 05, pp. 322-327, 2018.
[14] A. Manekar, and P. Gera, "Studying Cloud as Iaas for Big Data Analytics: Opportunity, Challenges," *International Journal of Engineering and Technology*, vol. 7, no. 2.7, pp. 909-912, 2018, doi: 10.14419/ijet.v7i2.7.11094.
[15] A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, "Providing Healthcare-as-a-Service using Fuzzy Rule Based Big Data Analytics in Cloud Computing," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1605-1618, 2018, doi: 10.1109/JBHI.2018.2799198.
[16] C. Lin, P. Wang, H. Song, Y. Zhou, Q. Liu, and G. Wu, "A Differential Privacy Protection Scheme for Sensitive Big Data in Body Sensor Network," *Annals of Telecommunications*, vol. 71, no. 9, pp. 465-475, 2016, doi: 10.1007/s12243-016-0498-7.
[17] Y. Yang, X. Zhen, W. Guo, X. Liu, and V. Chang, "Privacy-Preserving Fusion of IoT and Big Data for E-Health," *Future Generation Computer Systems*, vol. 86, pp. 1437-1455, 2018, doi: 10.1016/j.future.2018.01.003.
[18] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of Big Data Privacy," *IEEE Access*, vol. 4, pp. 1821-1834, 2016, doi: 10.1109/ACCESS.2016.2558446.
[19] P. Quinn, and L. Quinn, "Big Genetic Data and Its Big Data Protection Challenges," *Computer law & Security Review*, vol. 34, no. 5, pp. 1000-1018, 2018, doi: 10.1016/j.clsr.2018.05.028.
[20] H. Wang, "Anonymous Data Sharing Scheme in Public Cloud and Its Application in E-Health Record," *IEEE Access*, vol. 6, pp. 27818-27826, 2018, doi: 10.1109/ACCESS.2018.2838095.
[21] P. Li, S. Guo, T. Miyazaki, M. Xie, J. Hu, and W. Zhuang, "Privacy-Preserving Access to Big Data in the Cloud," *IEEE Cloud Computing*, vol. 3, no. 5, pp. 34-42, 2016, doi: 10.1109/MCC.2016.107.
[22] G. Pradeepini, G. Pradeepa, B. Tejanagasri, and S. H. Gorrepati, "Data Classification and Personal Care Management System by Machine Learning Approach," *International Journal of Engineering and Technology*, vol. 7, no. 2.32, pp. 219-223, 2018, doi: 10.14419/ijet.v7i2.32.15571.
[23] H. Dai, "Research on SVM Improved Algorithm for Large Data Classification," *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, Shanghai, China, 2018, pp. 181-185, doi: 10.1109/ICBDA.2018.8367673
[24] B. T. Pham, and I. Prakash, "A Novel Hybrid Model of Bagging-Based Naïve Bayes Trees for Landslide Susceptibility Assessment," *Bulletin of Engineering Geology and the Environment*, vol. 78, pp. 1911-1925, 2017, doi: 10.1007/s10064-017-1202-5
[25] S. Ataollah, T. B. Dieu, P. Binh, S. Karim, C. Kamran, K. Ataollah, S. Himan, and R. Inge, "Shallow Landslide Susceptibility Assessment using a Novel Hybrid Intelligence Approach," *Environmental Earth Sciences*, vol. 76, no. 60, 2017.
[26] B. Feizizadeh, M. S. Roodposhti, T. Blaschke, and J. Aryal, "Comparing GIS-Based Support Vector Machine Kernel Functions for Landslide Susceptibility Mapping," *Arabian Journal of Geosciences*, vol. 10, no. 5, pp. 122-135, 2017, doi: 10.1007/s12517-017-2918-z

[27] B. T. Pham, B. Pradhan, T. B. Dieu, I. Prakash, and M. B. Dholakia, "A Comparative Study of Different Machine Learning Methods for Landslide Susceptibility Assessment: A Case Study of Uttarakhand Area (India)," *Environmental Modelling & Software*, vol. 84, pp. 240-250, 2016, doi: 10.1016/j.envsoft.2016.07.005
[28] S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A. W. Nasir, V. Vijayakumar, and N. Chilamkurti, "Random Forest for Big Data Classification on the Internet of Things using Optimal Features," *International Journal of Machine Learning and Cybernetics volume*, vol. 10, pp. 2609-2618, 2019.
[29] K. S. Sree Ranjini, and S. Murugan, "Memory-Based Hybrid Dragonfly Algorithm for Numerical Optimization Problems," *Expert Systems with Applications*, vol. 83, pp. 63-78, 2017, doi: 10.1016/j.eswa.2017.04.033
[30] A. Chaudhary, S. Kolhe, and R. Kamal, "An Improved Random Forest Classifier for Multi-Class Classification," *Information Processing in Agriculture*, vol. 3, no. 4, pp. 215-222, 2016, doi: 10.1016/j.inpa.2016.08.002

BIOGRAPHIES OF AUTHORS

Snehalata K. Funde, Research Scholar, Koneru Lakshmaiah Education Foundation, Guntur, India. Snehalata received her Bachelor of Engineering and Master of Engineering degrees in Computer Science from the Savitribai Phule University, Pune, India. She has authored more than 10 articles in various domains like data mining, data security, and networking.

Gandharba Swain, Professor, Koneru Lakshmaiah Education Foundation, Guntur, India**.** Gandharba Swain received his MCA degree from University College of Engineering, Burla, M. Tech degree from NIT, Rourkela, and a Ph.D. degree from SOA University, Bhubaneswar. His research interests are blockchain, cryptography, steganography, and watermarking. He has authored more than 70 research articles and 2 books.