Acta Univ. Sapientiae Informatica **13,** 1 (2021) 1–20

DOI: 10.2478/ausi-2021-0001

Differential privacy based classification model for mining medical data stream using adaptive random forest

Hayder K. FATLAWI
ELTE University, Budapest, Hungary
University of Kufa, Najaf, Iraq
email: hayder@inf.elte.hu

Attila KISS
J. Selye University, Komarno, Slovakia
email: kissae@ujs.sk

Abstract. Most typical data mining techniques are developed based on training the batch data which makes the task of mining the data stream represent a significant challenge. On the other hand, providing a mechanism to perform data mining operations without revealing the patient's identity has increasing importance in the data mining field. In this work, a classification model with differential privacy is proposed for mining the medical data stream using Adaptive Random Forest (ARF). The experimental results of applying the proposed model on four medical datasets show that ARF mostly has a more stable performance over the other six techniques.

1 Introduction

A series of researches and projects in medical science, and information technology (IT) are starting a relationship between the healthcare industry and the IT industry that rapidly leads to a better and interactive relation among patients, their doctors, and health institutions. Data mining has a significant

Computing Classification System 1998: H.2.8, I.2.1
Mathematics Subject Classification 2010: 68P25, 97R40
Key words and phrases: ensemble methods, bagging, privacy-preserving protocol

role in medical data processing and analysis that mostly aims to predict the possibility of diseases or diagnose them [11]. One of the most remarkable challenges facing data mining is privacy preservation.

Privacy is an important component of medical data processing, as many health institutions refrain from providing this data to the public, due to the fear of compromising patient privacy. Therefore, providing a mechanism to carry out data mining operations, without revealing the patient's identity has recently taken place in the interest of researchers.

1.1 Problem statement

Privacy can be provided using many techniques that aim mostly to make a data modification to hide the identity of the objects in data and enable performing the mining operations on the data stream. This modification may destroy the distribution of the data, hence, the effectiveness of data will weaken for data mining techniques. Therefore, the combination of privacy and utility of the data for the mining process represents an interesting challenge. On the other hand, stream data mining techniques are characterized by fast response and ability to adapt to change in data distribution, while privacy-preserving techniques can cause delays in response time or/and difficulty in detecting the drift, which can lead to failure to adapt the mining model properly.

Also, the differential privacy-preserving technique performs data modification in which the average of added noise values for an attribute equal to zero, and that keeps the overall distribution of the data of this attribute. On the other hand, with a data stream, this can not be applicable because only some data instances in a specific time moment are available, which represents another challenge.

Therefore, the mining stream privacy-preserving model should satisfy the following conditions:

1. Data modification should be performed in which the presence or absence of any data element doesn't affect the statistics of the query. This condition aims to make any attacker can't ensure if any identity contributes to the data or not.
2. The modification should preserve the distribution changes in the stream samples to avoid decreasing classification accuracy.
3. Modification time should be fast as much as possible to avoid the response delay of the stream mining technique.

This work aims to design and implements a data stream classification model that satisfies these conditions. It should be capable of building a classifier

based on modified data to maintain privacy with minimal impact on response time and classification accuracy.

1.2 Related works

Chaudhuri et al. [5] addressed the tradeoff between privacy and learnability by focusing on privacy-preserving logistic regression. Their work involved disturbing the classifier with noise proportional to the sensitivity. They claimed that their technique didn't depend on the sensitivity of the function, and can be extended to a class of convex loss functions. Kadampur et al. [12] applied noise addition after building the decision tree from data in which for each path from the root to a leaf the noise values were added to the attributes of that path. Although the capability of handling categorical and numerical attributes, the classification accuracy was degraded after applying their model with three datasets.

Dwork et al. [7] constructed a privacy-preserving synopses using boosting for a set of queries over an input database, their algorithm obtains a synopsis that is good for all of these sets in which the privacy is guaranteed for the rows of the database while boosting is performed on the queries. They also provided synopsis generators for arbitrary sets of arbitrary low sensitivity queries. Vaidya et al. [20] utilized a random decision tree and random encryption to develop a distributed data mining framework with privacy-preserving. Their model had slower performance compared to a non privacy-preserving version though the accuracy exactly the same.

Two approaches from a combination of quasi-identifier and sensitive attribute (equal and unequal) were proposed by Bhaladhare et al. [3]. To minimize information loss that happened as a result of applying privacy-preserving, their model utilized systematic clustering for clusters generation. Although the loss of information and the execution time was better compared with Greedy k-member and Systematic clustering algorithms, their model had a moderate level of data utility. Homomorphic encryption scheme with cloud-aided association rule mining proposed by Li et al. [14] to achieve privacy-preserving with frequent itemset mining. According to their experiment results using many data sets, the model had fewer information leaks but higher computational time. Wang et al. [21] proposed a randomized response based approach for privacy-preserving in data collection. The implementation of their approach using data of patients showed less utility loss than the standard Laplace approach.

A distributed framework for preserving privacy using clustering in Hadoop was proposed by Nayahi et al. [16]. It used Hadoop Distributed File System and tried to overcome some attacks such as similarity attacks. The computational time of their model increased as the number of clusters increased, and they claimed that their algorithms were highly scalable with the size of the data set. Zhang et al. [22] used two mechanisms of noise: Laplace and exponential for providing privacy. They utilized lower noise sensitivity to avoid a high impact on split point choosing. They applied the proposed model on only one dataset, and the results showed more stability in classification accuracy compared with three other algorithms.

Beck et al. [19] proposed a data analytics system for privacy-preserving of a data stream, it provided zero-knowledge privacy guarantee for users, a data analysts interface to explore the output accuracy with the query execution budget, and a close real-time stream processing based on a scalable distributed architecture. Manikandanet al. [15] utilized a code-based threshold scheme with fuzzy c-means clustering for creating distributed privacy-preserving.

Table 1 summarizes the characteristics of those related works which have been mentioned in this section. Most of the related works mentioned in Table 1 were involving classification tasks and only one of them was working with stream data. This points to the lack of research works in privacy preservation for stream data mining. Also, those classification works were mostly lacking in utilizing ensemble classifiers which have a preferable performance with real-world datasets.

This paper is an extension to our paper [8], the extension utilizes the robustness of Adaptive Random Forest (ARF) ensemble classifier against small changes in the distribution of stream data, and build a classification model with the ability of privacy-preserving using Laplace distribution instead of normal distribution. The extension includes mentioning and analysis for additional related works, also the evaluation of the proposed model including one addition algorithm, two new datasets, a different range of noise values that added to the data, and a new comparison for distribution changes and adaptive window sizes. The implementation of the proposed model in this extension produces 364 experiments and confirms the better performance of ARF.

Article	Data Mining Tech.	Dataset	Privacy Preserving Technique	Advantages	Disadvantages
Chaudhuri et al. [5]	Logistic Regression	Artificial Dataset	Differential Privacy	Sensitivity Independent	Only Simulation Results
Kadampur et al. [12]	Decision Tree	Boston Housing Price,Census Income,Car Evaluation	Noise Addition	Handle Categorical and Numerical Data Types	Less Accuracy than Original Classifier
Dwork et al. [7]	Boosting	-	Differential Privacy	Stronger Bounds on Expected Privacy Loss	No Experiment on Real Datasets
Vaidya et al. [20]	Random Decision Tree	Mushroom, Nursery, Image Segmentation, and Car	Random Encryption	Fast Distributed Mining with Same Accuracy	Slower than Non Privacy-preserving Version
Bhaladhare et al. [3]	Systematic Clustering	Benchmark Adult	Combination of Quasi-identifier and Sensitive Attribute	Lesser Information Loss	Moderate Level of Data Utility
Li et al. [14]	Frequent Itemset	Retail and Pumsb Datasets	Vertically Partitioned Databases	Leak Less Information	Slower than Algorithms with Low Privacy Levels.
Wang et al. [21]	Data Collection	YesiWell	Randomized Response	Fewer Utility Loss with High Sensitivity of Functions	Depending on Only One Dataset
Nayahi et al. [16]	J48 , Naive Bayes , K-NN	Benchmark Adult d	K-anonymization	Scalability on Increasing Dataset Size	Time Increasing when Number of Clusters Increasing
Zhang et al. [22]	Decision Tree	Census Income	Laplace and Exponential Noise	More Stable Accuracy	Depending on Only One Dataset
Beck et al. [19]	Sampling	NYC Taxi Ride,Household Electricity Consumption	Randomized Response	Distributed Real-time Stream Processing	Accuracy Loss doesn't Always Decrease when Second Randomization Parameter Increases
Manikandan et al. [15]	Fuzzy C-Means	Plant Cell Signaling	Code Based Technique with Threshold Estimation	Less Number of Iterations and No Cross Trust is Required	Focus Only on Efficiency

Table 1: Comparison of some research works on privacy-preserving data mining

2 Basic concepts in stream data mining

2.1 Data stream constraints

Unlike with batch data, stream data faces many constraints as follow: (1) infinite arrival of data samples make storing them impossible, (2) the fast arrival of data samples requires dealing with each sample in real-time, (3) the possibility of changing items' distribution overtime in which the old data would be useless for the current status. Generally, the perfect classification model should produce maximum accuracy in the fastest time and minimum computational resources [2].

2.2 Concept drift

Concept drift refers to that the data is being gathered may change from time to time, every time according to some minimum persistence. Changes may occur during the time in which the old training examples become irrelevant to the current state, and the learning system should forget such kind of information. There are two important issues related to the change: causes of change and the rate of change [9].

2.3 Adaptive sliding window (ADWIN)

It is an estimation technique that aims at detecting the change in a data stream based on a sliding window with adaptive size. It has a qualified and significant method for tracking the average of bits in the stream. In this technique, the length of windows is not updated as long as the average value inside the window doesn't change [9].

2.4 Hoeffding tree

Hoeffding Tree or Very Fast Decision Tree (VFDT) is a variation from the typical decision tree designed for stream data. The learning of these techniques depends on replacing leaves of the tree with decision nodes. Each terminal node (leaf) in the tree stores enough information statistics about features values that are used by a heuristic function to perform a splitting test. After reaching a new data instance, it transfers starting from the root until reaching a specific leaf node. At this point, the statistics information then is evaluated and a new decision node may be created based on this evaluation [9]. It is very popular to

utilize VFDT as a base learner for the ensemble classification model, thereby, the ensemble techniques in this work used VFDT as well.

VFDT depends on the concept of Hoeffding Bound [9] which states that the probability of the difference between the expected value and the actual value of the mean of data elements to be more than ϵ value shouldn't exceed a specific small value as follows: let $F_1, F_2, \ldots, F_n$ be an independent random variable and each F_i is bounded in which

$$P\left(F_i \in R = [x_i, y_i]\right) = 1. \tag{1}$$

Let

$$H = \frac{1}{n}\sum_{i=1}^{n} F_i$$

with expected value $E(H)$. Then for any $\epsilon > 0$,

$$P[H - E[H] > \epsilon] \leq e^{-\frac{2n^2\epsilon^2}{R^2}}. \tag{2}$$

2.5 Adaptive random forest

Models based on a single classifier have some weakness points, such as model instability which means any slight changes in data may make a change in the structure of the tree classifier. To overcome that, ensemble methods have been developed which combine many weak classifiers [13, 1]. Ensembles have more power predictive performance than a single tree, so they became general techniques for both classification tasks and numeric prediction [17, 13]. The methodology of an ensemble model is to combine a set of single models, each one tries to solve the same original task, aiming to obtain a better integrated global model [10]. Two points should be taken into account when using ensembles: (i) the size of an ensemble (ii) the mechanism of combination among the results of trees [23]. Many techniques are developed for ensemble models such as bagging, boosting, and stacking. Bagging combines the decisions of multiple trees by using the voting concept for binary (and multi) class predictive tasks, and for a numerical predictive task, bagging calculates the average. A popular example of bagging techniques is the random forest [1].

Ensemble modeling aims at building a strong accumulative classifier from many weak classifiers. Adaptive Random Forest is a variation from the typical random forest algorithm for data stream mining tasks. The main idea is to utilize Hoeffding trees, which have the ability to adapt to distribution changes, as the base classifier for the bagging ensemble method [2]. For detecting the

change in a data stream, ADWIN is used in these techniques. It depends on Online Bagging as a resampling method and a drift monitor for change detection per each tree [2].

2.6 Differential privacy

Differential privacy aims at learning information about the whole data while preserving the privacy of each data sample. The differential privacy model assumes that although the availability of the knowledge about all data records except one, the adversary is not be able to extract the information of that record. It can be resistant to background attack in comparison with other privacy models, also, privacy guarantee of differential privacy is provable [24].

In general, the system with differential privacy should have the same performance without any consideration for to presence or absence of any data sample, and this can be performed by keeping the probability distribution of data [7]. According to [5], differential privacy can be provided using a randomized mechanism RM if for all databases $\mathsf{DB1}$ and $\mathsf{DB2}$ that differ by one element for any t,

$$\frac{\mathsf{P}[\mathsf{RM}(\mathsf{DB1}) = \mathsf{t}]}{\mathsf{P}[\mathsf{RM}(\mathsf{DB2}) = \mathsf{t}]} \leq e^{\epsilon} \tag{3}$$

The privacy guarantee level of the differential privacy model is controlled by the parameter ϵ which represents the privacy budget. Sensitivity is another aspect related to differential privacy, it indicates the required amount of perturbation for this mechanism by calibrating the volume of noise. There are two types of sensitivity used in differential privacy: (i) global sensitivity represents the largest value for the difference between results of the query on different related datasets, (ii) local sensitivity concern with calibrating the difference between query results based on records. Queries with relatively low values are preferable with global sensitivity [24].

3 Methodology

The main aim of this work is to design and implement a classification model for stream data based on adaptive random forest, including differential privacy. Thereby, there are two main stages; the first one is to apply some of the preprocessing procedures to prepare the medical data for the mining task. The second stage is to build an ensemble classifier which includes many very

fast decision trees, and finally compare the performance based on streaming real batch datasets. Figure 1 illustrates all the steps of our work.

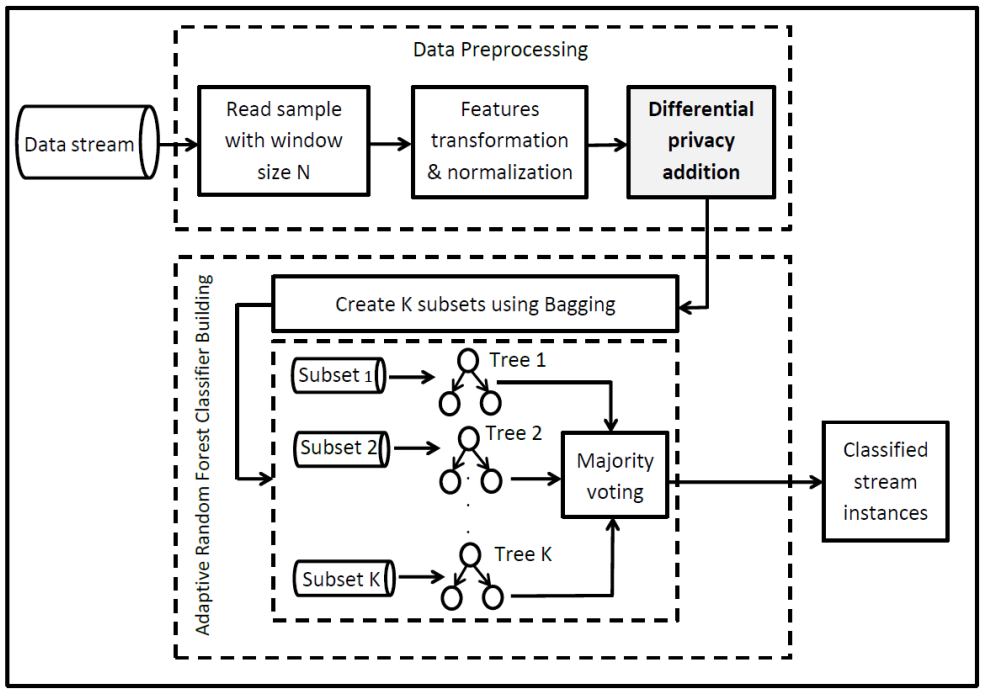

Figure 1: The classification model with ϵ-differential privacy for medical data stream

3.1 Stage one: data preprocessing

The first stage concerns with preparing and generating a new dataset from the original one by using features transformation, normalization, and the white noise concept. This stage can be described as following steps:

3.1.1 Categorical to numerical transformation

To add the noise values to the data, features should be in a numeric form. So, in this step, every Categorical (Textual values) was converted to numerical values. For binary features values like (Yes, No), the simplest coding is used and the values became (1,0). For multiple values (more than two values), frequency of

each distinct value (1..N) for each feature was calculated. After that coding was used in which the most frequent distinct textual value converted to N and least frequent converted to 1.

3.1.2 Data normalization

The range of feature values can be different, such as age has values between 1 - 150 while the yearly income can be between (1-10000000). For that, we need to apply normalization to prevent any dominant from one of the features during statistical calculation that performed during classifier building. The new range for all feature values were between -1 and 1 in which for each feature f:

$$f(i) = 2\frac{f(i) - minf}{maxf - minf} - 1 \tag{4}$$

3.1.3 ϵ-Differential noise generation

For each data set, noise value was added in which the mean of those values for each feature is zero. The standard deviation (STD) represents the intensity of the noise and the gradual increase of it will be used in the proposed model to investigate the suitable value for the input data. The random values should satisfy the condition of the random mechanism Eq. (3). To obtain these noise values, Laplace Mechanism as presented in [24] was utilized in which the values were generated from the Laplace distribution, which has zero center and scale q. Large q value produces a higher noise value z as the following:

$$Lab(z) = \frac{1}{2q} exp\left(\frac{-|z|}{q}\right). \tag{5}$$

3.2 Stage two: building ensemble model

In this stage, we utilized online Bagging of K base classifier [18], each base classifier built using Hoeffding Tree as presented by [6] . The stage started with setting the size of ARF ensemble model, then number of data subsets produced from resampling the data sample that resulted from the previous stage. The number of subsets was equal to the number of base classifiers, each subset was used to train a Hoeffding Tree classifier, and finally, voting among all base classifiers was used to classify each data element.

3.2.1 Initializing the size of ensemble model

While Online Bagging was used for ensemble model building, the size of this model i.e. the number of base classifiers was a user-defined parameter that needs its value before the building operation started. The importance of this parameter comes from it represents a stopping condition for each learning step, which is related to the complexity of required computational resources. The value of this parameter in the proposed model prefers to be in low range value (around 10 base classifier) for the following reasons:

1. Stream mining classifier is expected to have a fast response in learning and classifying process as a stream element reach continuously.
2. The number of data rows in each data sample inside the current window is relatively low, thereby, the resampling step in Online Bagging doesn't need for large ensemble model for preserving the diversity of each base classifier.

3.2.2 Ensemble model building

In online Bagging, any new data sample was chosen according to a Poisson(1) distribution. The classification decision of the ensemble bagging model is based on the voting of all K base classifiers with equal weight for all of them. It gives every new data example an initial weight w=1, then it is passed to the first weak learner. If this data example is misclassified, it's weight is increased before passing it to the next weak learner. The base learner in our comparison was Hoeffding Tree classifier and the size of the ensemble that used was ten learners. Adaptive Random forest was built depending on [4] which utilized Online Bagging's resampling method but the difference was in the adaptive method.

3.2.3 Hoeffding tree classifier building

It includes two types of nodes: internal (or decision) and terminal (or leaf) nodes. Each terminal node in the tree stores enough statistical information about features values. This information is used by a heuristic function to perform a splitting test. After reaching a new data instance, it transfers starting from the first node (root) until reaching a specific leaf node. In this point, if the class value of new instance isn't seen before, the instance then is classified according to the majority class of the current leaf node, otherwise, the statistics information is evaluated and a new decision node may be created based on this evaluation.

The evaluation includes computing the gain for all features in all possible split points. For each split point, the impurity of class distribution of the current node and the possible child nodes will be computed using Entropy, according to the following equation, for node a:

$$\mathtt{Entropy}(a) = -\sum_{i=0}^{c-1} p\,(i|a)\,\log_2 p\,(i|a)\,, \tag{6}$$

where c refers to the number of classes. The difference between the Entropy of the current node and the average of Entropy of its possible child nodes after splitting represents the gain of that splitting operation. A splitting that produces a more homogeneous class distribution i.e. higher gain is preferable. Hoeffding bound computes using Eq. (2), and if the difference between the highest two features is more than Hoeffding bound value, the current leaf node will replace by an internal decision node depending on the highest feature, also, for each split branch of this new node, a new empty leaf node will be added. Algorithm 1 summarizes all the steps of the proposed model.

4 Implementation and experimental results

4.1 Data analysis platform

In this work, three major tools were utilized to perform the comparison; Waikato Environment for Knowledge Analysis (Weka), Massive Online Analysis (MOA), and Sklearn. Weka Platform is open-source software for data analysis tasks including Classification, Clustering, and Association Rules. It is developed by the University of Waikato using Java programming language. It was utilized in this work for preprocessing operations (Transformation and Normalization).

MOA Platform is an improvement for the Weka platform for the mining data stream. It provides many popular mining techniques, stream generator, and concept drift detection techniques, in our comparison, it performed the data streaming and implementation of classification techniques. Sklearn is a python free library for machine learning tasks. It contains many classification techniques such as random forest and boosting. Sklearn was used in this work for adding white noise values to data.

Algorithm 1 Stream data classification model with ϵ-differential privacy

procedure ARF
    St = current stream samples in ADWIN window
    From St Create K data subsets using Bagging resampling
    for each subset S **do**
        for each feature f in S **do**
            if IsCatogorical(f) = True **then**
                for each distinct value d in f **do**
                    f(d)= frequency(d)
                end for
            end if
            Apply normalization on S according to Eq. (4)
        end for
        for each data instance dt in S **do**
            Generate random value R for differential privacy ▷ Based on Eq. (3),(5)
            dt = dt+R
            Trace Hoeffiding tree reaching to a specific terminal node tn
            if Class value of dt != ? **then**
                dt(y)= major class of tn
            else
                for each feature f in tn **do**
                    G1= Class impurity in tn ▷ Based on Eq. (6)
                    G2= Class impurity in tn 's possible child nodes ▷ Based on Eq. (6)
                    G= G1-G2
                end for
                Rank every features based on its gain
                Choose two features bf1,bf2 with highest gain
                Compute HB= Hoeffiding Bound based on Eq. (2)
                if G(bf1)-G(bf2) > HB **then**
                    Replace tn with a decision node based on split test of bf1
                    Add new terminal nodes for each possible split value
                end if
            end if
        end for
    end for
    Return ARF classifier, Classified sample instances
end procedure

4.2 Applying adaptive random forest with differential privacy

This step includes applying ensemble classifier against gradually increasing in differential privacy strength of data by using white noise. For this task, the Weka platform is used to perform two preprocessing steps; features transformation and normalization. Then MOA is used to convert batch datasets to a data stream, then to train the classifier based on that stream. Four measurements are used for evaluating the performance of the classification techniques; mean of correctly classified instances, mean of F1 score, mean of precision, and mean of Recall. Figure 2 and Table 2 clarify the performance of the proposed model using different Standard Deviation STD randomize values for Differential Privacy.

Diff. Noise STD	0.000	0.001	0.002	0.003	0.004	0.005	0.01	0.015	0.02	0.025	0.05	0.075	0.1
EEG State	98.75	98.28	98.17	98.29	98.11	98.08	98.03	98	97.90	98.06	98.06	98.09	97.84
Skin Seg.	100.00	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99	100.00	99.99	99.99	99.99
MIT-BIH	90.93	82.72	82.71	82.71	82.70	82.71	82.71	82.70	82.70	82.71	82.71	82.72	82.70
Breast Cancer	99.35	99.35	99.35	99.35	99.35	99.35	99.35	99.35	99.35	99.35	99.35	99.35	99.35

Table 2: Accuracy of ARF with different STD values

Figure 2: Accuracy of ARF with range of STD values for differential privacy

In both Table 2 and Figure 2 , we can observe the following:

1. Ability of ARF to preserve the accuracy of classification after adding the noise values with the first two and the fourth datasets.
2. Stability of ARF with different values of additional noise with the first two and the fourth datasets.
3. There was a decrease in the classification accuracy with the third dataset after adding the minimum magnitude of the noise, however, ARF recovered its stability with the rest of the range's values.

The main difference between the first two and the fourth dataset aside, and the third dataset from another side is that the number of features in the third dataset is more, this leads to a question that if the high dimensionality can affect the utility of ARF after adding the differential privacy.

The preference of the proposed model using ARF compared with many other techniques can be observed in Table 3 and Figure 3, however, there was a close performance between ARF and Ozabagging. The similarity of those two techniques that are both of them is a bagging ensemble classifier, a question arises if the strength of ARF with privacy-preserving in the medical data stream can be generalized to other bagging techniques. Also, we can observe that Naive Bayesian had unstable performance, in which it had the worst performance in most cases. All results in Table 3 were using the minimum STD value for differential privacy.

Technique	Heoffman	ARF	OzaBagg	OzaBoost	K-NN	N. Baysain	Random Hoeffman
EEG State	72.56	98.28	93.33	85.56	88.4	48.42	65.48
Skin Seg.	99.94	99.99	99.97	79.06	99.97	95.29	99.93
MIT-BIH	82.7	82.72	82.72	87.87	82.67	14.61	82.70
Breast Cancer	99.31	99.35	99.35	99.19	99.35	94.33	99.3

Table 3: Accuracy comparison of classification algorithms based on streaming four medical datasets

Other interesting findings from the experimental results are illustrated in Figure 4 and Figure 5. We can observe that the number of drifts i.e change in the distribution of data streams for all features in Skin Seg. and EEG Eye datasets was reduced significantly after adding differential privacy. As a result of this reduction, the size of ADWIN window which illustrated in Figure 6 and Figure 7 was maximized to contain all stream elements in EEG state and Skin Seg. datasets. This smoothness of the data stream leads to reduce the number of changes in ARF model to adapt to change in distribution, thereby,

the computational time of ARF has been reduced, and that could overcome the addition time of differential privacy step.

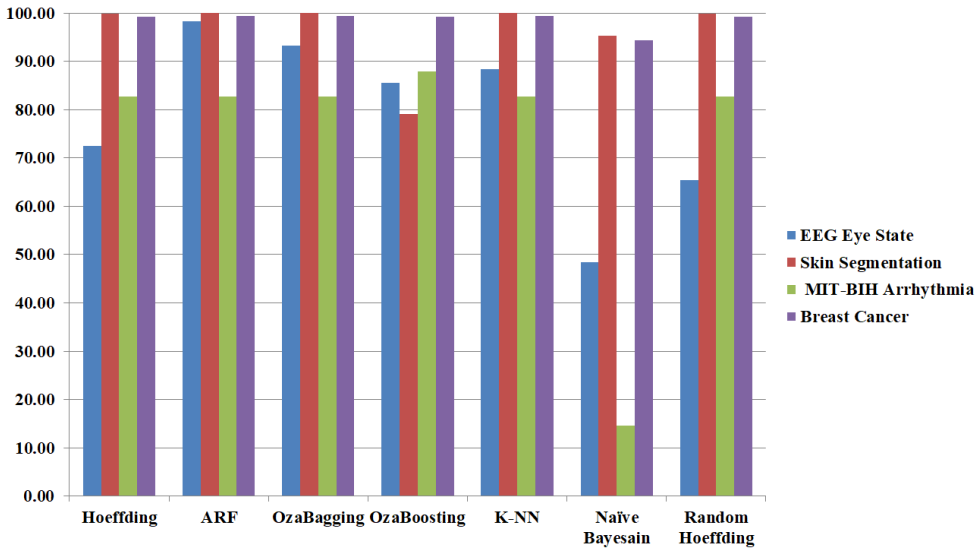

Figure 3: Comparison of the proposed model accuracy with other six techniques

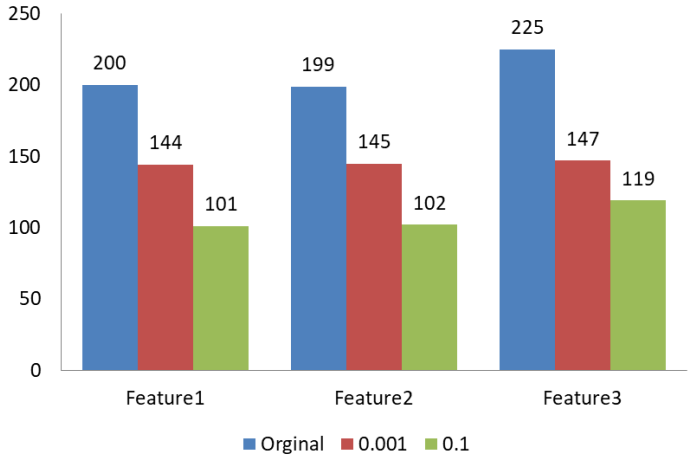

Figure 4: Comparison of drifts in skin sig. dataset with two STD values

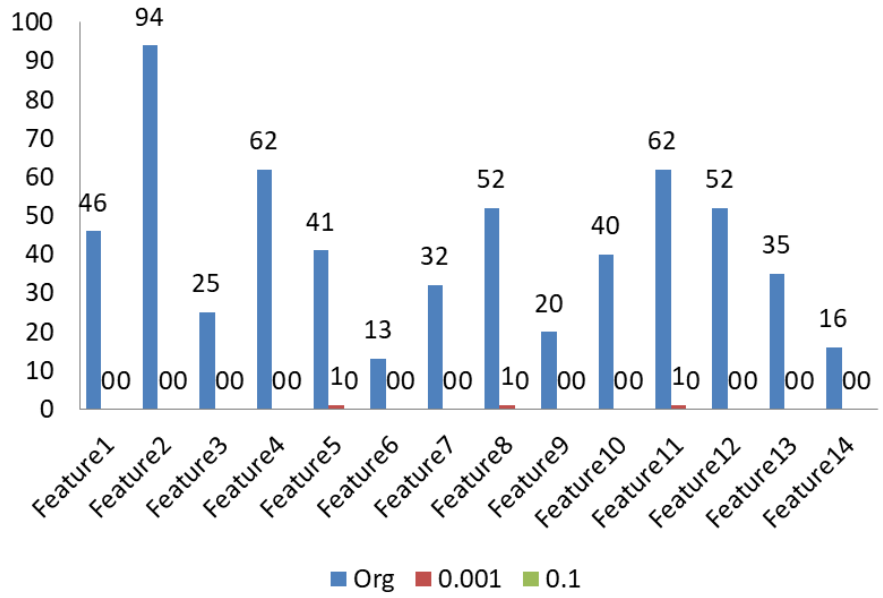

Figure 5: Comparison of drifts in EEG state dataset with two STD values

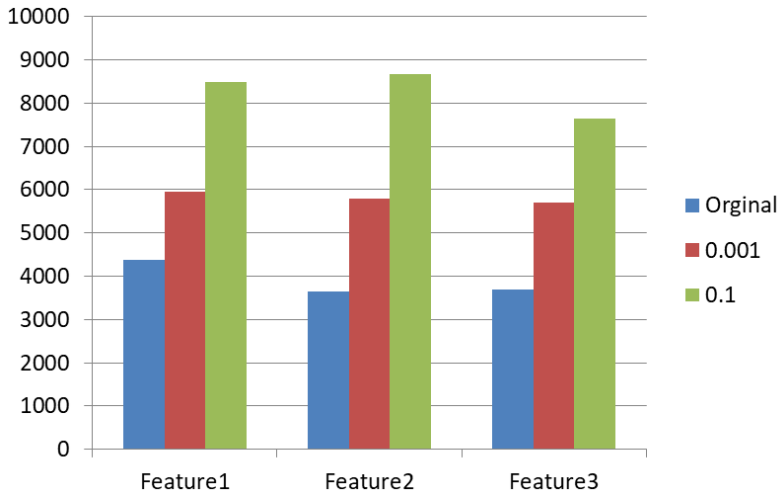

Figure 6: Comparison of ADWIN size in skin sigm. dataset with two STD values

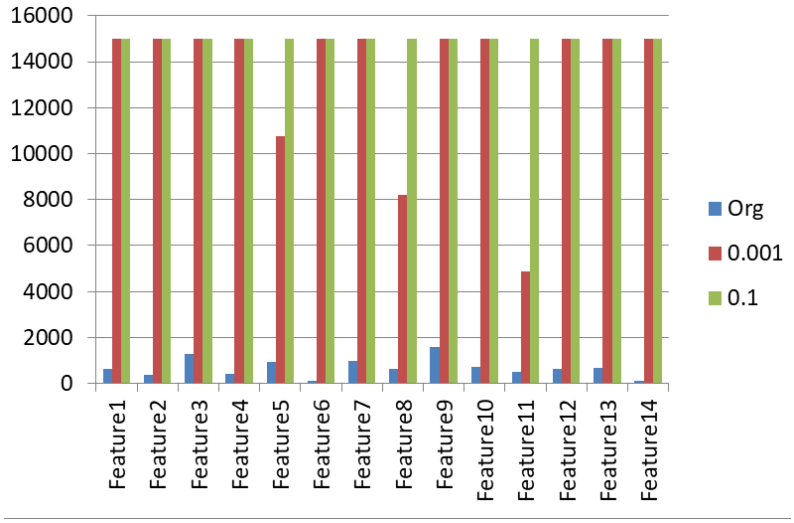

Figure 7: Comparison of ADWIN size in EEG state dataset with two STD values

5 Conclusion

In this work, a privacy-preserving classification model for the mining data stream was proposed. It utilized the robustness of Adaptive Random Forest classifier to handle the randomized values that added to the original data stream using differential privacy. The proposed model had the best accuracy compared with the other six techniques applied on four real medical datasets, OzaBagging also has notable performance, and both techniques are bagging ensemble methods. Also, the number of drifts in the distribution of data streams was reduced significantly after adding differential privacy, as a result, that the size of ADWIN window was maximized. These results obtained by applying 364 experiments using a gradual increase of STD of randomizing values for proving the differential privacy.

Acknowledgments

The project was supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

References

[1] A. Al-Fatlawi, H. Fatlawi, S. H. Ling Recognition physical activities with optimal number of wearable sensors using data mining algorithms and deep belief network, *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* Seogwipo, South Korea, 2017, pp. 2871–2874. ⇒7

[2] B. Babenko, MH. Yang, S. Belongie, A family of online boosting algorithms, *2009 IEEE 12th international conference on computer vision workshops, ICCV workshops* Kyoto, Japan, 2009, pp. 1346–1353. ⇒6, 7, 8

[3] P. R. Bhaladhare, D. C. Jinwala, Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model, Journal of information science and engineering, **32** (2016) 63–78. ⇒3, 5

[4] A. Bifet, G. Holmes, B. Pfahringer, G. Bernhard, Improving adaptive bagging methods for evolving data streams, *Asian conference on machine learning* Nanjing, China, 2009, pp. 23–37. ⇒11

[5] K. Chaudhuri, C. Monteleoni, Privacy-preserving logistic regression, *Advances in neural information processing systems* Vancouver, Canada, 2009, pp. 289–296. ⇒3, 5, 8

[6] P. Domingos, G. Hulten, Mining high-speed data streams, *KDD00: the second annual international conference on knowledge discovery in data* Boston Massachusetts, USA, 2000, pp. 71–80. ⇒10

[7] C. Dwork, G. N. Rothblum, S. Vadhan, Boosting and differential privacy, *2010 IEEE 51st annual symposium on foundations of computer science* Las Vegas, Nevada USA, 2010, pp. 51–60. ⇒3, 5, 8

[8] H. Fatlawi, A. Kiss, On Robustness of Adaptive Random Forest Classifier on Biomedical Data Stream, *Asian Conference on Intelligent Information and Database Systems (ACIIDS 2020)*, *Lecture notes in computer science* Springer, **12033** (2020) 332-344. ⇒4

[9] J. Gama, *Knowledge discovery from data streams*, The CRC Press, 2010. ⇒6, 7

[10] G. Giovanni, J. F. Elder, *Ensemble methods in data mining: improving accuracy through combining predictions*, The Synthesis lectures on data mining and knowledge discovery Morgan & Claypool Publishers, 2010. ⇒7

[11] G. Hulten, L. Spencer, P. Domingos, Mining time-changing data streams, *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining* San Francisco California, USA, 2001, pp. 97–106. ⇒2

[12] M. A. Kadampur, D.V.L.N Somayajulu, A noise addition scheme in decision tree for privacy preserving data mining, Journal of computing **2,** 1 (2010) 137–144. ⇒3, 5

[13] M. Kuhn, K. Johnson, *Applied predictive modeling*, Springer, 2013. ⇒7

[14] L. Li, R. Lu, K. R. Choo, A. Datta,J. Shao, Privacy-preserving-outsourced association rule mining on vertically partitioned databases, IEEE transactions on information forensics and security, **11,** 8 (2016) 1847–1861. ⇒3, 5

[15] V. Manikandan, V. Porkodi, A. S. Mohammed, M. Sivaram, Privacy preserving data Mining using threshold based fuzzy C-Means clustering, ICTACT journal on soft computing, **9,** 1 (2018) 1820–1823. ⇒4, 5

[16] J. J. v. Nayahi, V. Kavitha, Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop, Future Generation Computer Systems, **74** (2017) 393–408. ⇒4, 5

[17] T. Ngo, *Data mining: practical machine learning tools and technique, by ian h. witten, eibe frank, mark a. hell*, The ACM SIGSOFT Software Engineering Notes, **36,** 5 2011. ⇒7

[18] N.C. Oza, Online bagging and boosting, *2005 IEEE international conference on systems, man and cybernetics* Waikoloa, HI, USA, 2005, pp. 2340–2345. ⇒10

[19] D. L. Quoc, M. Beck, P. Bhatotia, R. Chen, Christof Fetzer, Thorsten Strufe, PrivApprox: privacy-preserving stream analytics, *2017 annual technical conference (USENIX ATC '17)* Santa Clara, CA, USA 2017, pp. 659–672. ⇒4, 5

[20] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, D. Lorenzi, A random decision tree framework for privacy-preserving data mining, IEEE transactions on dependable and secure computing, **11,** 5 (2013) 399–411. ⇒3, 5

[21] Y. Wang, X. Wu, D. Hu, Using Randomized Response for Differential Privacy Preserving Data Collection, EDBT/ICDT Workshops, Bordeaux, France, **1558** (2016). ⇒3, 5

[22] L. Zhang, Y. Liu, R. Wang, X. Fu, Q. Lin, Efficient privacy-preserving classification construction model with differential privacy technology, Journal of systems engineering and electronics BIAI, **28,** 1 (2017) 170–178. ⇒4, 5

[23] Z. Zhou, *Ensemble methods: foundations and algorithms*, The CRC press, 2012. ⇒7

[24] T. Zhu, G. Li, W. Zhou, S. Y. Philip, *Differential privacy and applications*, Springer, 2017. ⇒8, 10

Received: January 23, 2021 • Revised: February 28, 2021