2-1-2024

EDSUCh: A robust ensemble data summarization method for effective medical diagnosis

Mohiuddin Ahmed
Edith Cowan University

A. N. M. B. Rashid
Edith Cowan University

10.1016/j.dcan.2022.07.007
Ahmed, M., & Rashid, A. N. M. B. (2024). EDSUCh: A robust ensemble data summarization method for effective medical diagnosis. Digital Communication and Networks, 10(1), 182-189. https://doi.org/10.1016/j.dcan.2022.07.007

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.keaipublishing.com/dcan

EDSUCh: A robust ensemble data summarization method for effective medical diagnosis

Mohiuddin Ahmed [*], A.N.M. Bazlur Rashid

School of Science, Edith Cowan University, Perth, Australia

ARTICLE INFO

Keywords:
Data summarization
Ensemble
Medical diagnosis
Sampling

ABSTRACT

Identifying rare patterns for medical diagnosis is a challenging task due to heterogeneity and the volume of data. Data summarization can create a concise version of the original data that can be used for effective diagnosis. In this paper, we propose an ensemble summarization method that combines clustering and sampling to create a summary of the original data to ensure the inclusion of rare patterns. To the best of our knowledge, there has been no such technique available to augment the performance of anomaly detection techniques and simultaneously increase the efficiency of medical diagnosis. The performance of popular anomaly detection algorithms increases significantly in terms of accuracy and computational complexity when the summaries are used. Therefore, the medical diagnosis becomes more effective, and our experimental results reflect that the combination of the proposed summarization scheme and all underlying algorithms used in this paper outperforms the most popular anomaly detection techniques.

1. Introduction

Because of the rapid advancement of computing and communication technology, the Internet of Things (IoT) has connected to incredible devices of diverse collections. The Internet of Health Things (IoHT) is an example of IoT in the health sector. IoT or IoHT generates a massive amount of data, often called Big Data [1,2]. While Big Data creates opportunities to explore and research innovations, it can create doors for effective medical analysis by identifying important biomarkers or identifying anomalous patterns in the data for developing new diagnoses. Big datasets can have rare or infrequent patterns, and detecting these patterns is computationally expensive [3–5]. A lack of analysis of these rare patterns in the medical domain may take a lot of lives. Hence, it is essential to offer a computationally inexpensive anomaly detection approach for effective medical diagnosis to identify rare patterns or diseases with higher detection accuracy. However, it is quite challenging to effectively analyze large datasets and identify anomalies with reduced computations. Therefore, a summary that is a concise version of the original data can help the data science community effectively analyze such large datasets in any domain. Detail on the summary and the critical analysis of existing data summarization techniques can be found in Ref. [6]. A summary of data that represents useful information from the original data with an appropriate ratio to all types of data can be used for anomaly detection, i.e., for effective medical diagnosis. In general, a data summarization can be performed using clustering or frequent itemsets. However, sampling techniques can be considered for data summarization because data sampling has been a proven method for compressing input data in many domains.

In literature, several ensemble data summarization approaches are studied using clustering, and sampling techniques, such as in Refs. [7–10]. While the clustering methods can group similar samples into one group, the sampling techniques can be used for selecting samples from the clustered outcomes. However, there is a challenge of how to choose the sampling size. This paper aims to investigate the ensemble data summarization approach for effective medical diagnosis based on the motivation of using clustering and sampling together. Therefore, in this paper, an ensemble data summarization method has been proposed by combining clustering and sampling techniques. From the experimental analysis of the medical data, it can be observed that the proposed ensemble approach can ensure the inclusion of rare patterns in the summary data. The proposed ensemble data summarization uses bootstrap sampling techniques, and Chernoff bound for computing the sampling size to reduce the loss of information in the summary data. Therefore, the proposed approach is termed as EDSUCh. Two clustering algorithms were applied separately and combined on four benchmark medical datasets before the sampling techniques were applied to produce

[*] Corresponding author.
E-mail addresses: m.ahmed.au@ieee.org (M. Ahmed), a.rashid@ecu.edu.au (A.N.M.B. Rashid).

https://doi.org/10.1016/j.dcan.2022.07.007
Received 10 April 2022; Received in revised form 11 July 2022; Accepted 17 July 2022
Available online 31 July 2022

the summary. The performance of the experimental results was evaluated based on four anomaly detection techniques, both on the original and the summary data. The application of the proposed EDSUCh requires fewer computations to create summary and anomaly detection compared to the anomaly detection on the original data.

Contribution of this study

The fundamental research question and associates sub-questions in this paper are:

- How can an ensemble data summarization method be applied to produce a summary from the original data, which includes interesting data patterns, such as the rare patterns for effective medical diagnosis? How can effectively the data summarization include rare anomalous samples in the summary data?
  - How can the clustering and sampling method be applied to produce the summary?
  - How can the summarization determine the appropriate size of the sampling and summary?
  - Can anomaly detection be performed both in the summary and original data?
  - Can the data summarization improve anomaly detection performance?
  - Can the data summarization reduce computations for anomaly detection?

The summary needs to include rare patterns for anomaly detection using data summarization. When the summary contains only a set of normal samples while the original data has both normal and anomalous samples, the anomaly detection from the summary can be useless. The distribution of anomalous samples in summary and the original data may vary, resulting in variations in anomaly detection performance. Therefore, the right combination of data summarization and anomaly detection can improve the detection performance. An ideal summary of data should contain rare patterns and also be concise. Hence, an appropriate sampling size is important for data summarization. While summary size has an impact on the computations for anomaly detection, the inclusion of appropriate rare anomalies can ensure improved detection performance. Anomaly detection from original data takes the detection time. On the other hand, the time required for creating the summary and anomaly detection should be considered using a summary. When summary data can ensure reduced computations and improved anomaly detection performance, it should be considered the effective summary for the underlying domain dataset.

Rest of the paper is organized as follows. Section 2 discusses the relevant works. Section 3 presents the proposed data summarization technique for effective medical diagnosis. Section 4 contains the experimental results and analysis. Section 5 discusses the findings and concludes the paper.

2. Relevant works

Data summarization is an effective and proven technique for creating a data summary representing the original dataset. In literature, several data summarization methods utilized the ensemble approaches. For example, the clustering results were used to boost classifiers' prediction performance for sensor data by Lavanya et al., in 2021. The effectiveness of the proposed ensemble approach was examined for feature selection techniques, including rough set and entropy, and validated using naive Bayes, *k*-nearest neighbor, support vector machine, and decision tree classifiers [7]. Ensemble techniques used for the text summarization by using the best features of existing text summarization methods, such as intra-sentence and inter-sentence cosine similarities [8]. Some other ensemble approaches are studied in the literature, such as in Refs. [11–14]. Ensemble of clustering is a process of aggregating different decisions obtained by many clustering algorithms. This can lead to overcoming individual clustering algorithms' shortcomings and improving the clustering performance. The clustering ensemble generally requires two stages: generation of cluster ensemble and a consensus function. In the first step, the ensemble determined the members to be added to the sample by ensuring the diversity for the cluster quality improvement. Different approaches, such as random subsampling and homogenous or heterogeneous ensemble methods, can be used to achieve the clustering outcome. In the second step, the final sample of the ensemble can be obtained by combining the results of individual clustering algorithms. Based on this motivation, a good ensemble approach can be studied using clustering and sampling techniques to create an efficient summary that can represent the original dataset appropriately for effective medical diagnosis. For the critical part of choosing the sampling size for a cluster sample or ensemble, the widely used Chernoff bound [15] method can be used, which was also utilized in a previous study successfully [9,10]. Only an appropriate data summary can improve the machine learning performance and decrease the computations.

3. Proposed ensemble data summarization method for effective medical diagnosis

The proposed ensemble data summarization Using Chernoff Bound (EDSUCh) for effective medical diagnosis is based on sampling. The motivation to integrate a sampling method in ensemble data summarization is based on the requirement to use original data samples in the summarized data, unlike other methods, such as clustering or frequent itemsets. Sampling has been a proven technique to compress the input data and has been effectively used in different management activities, including intrusion detection, traffic characterization, and anomaly detection [3–5,16]. The fundamental benefits of using a sampling method are reducing cost and speeding up the execution process over the complete enumeration [17]. There are different types of sampling methods available in practice, and the bootstrapping sampling method has been adopted in the proposed EDSUCh approach. Bootstrapping sampling is a classical method that assesses the variability of a sample statistic. It uses multiple samples with replacement from the observed dataset [18]. One of the fundamental challenges in the data summarization and sampling process is choosing the sampling size. An optimal sampling size can intelligently reduce the input data size and keep the data samples ratio per type of sample. To compute the sampling size for the summarized data, *Chernoff bound* [19] has been incorporated. Using *Chernoff bound*, it can be analytically shown that the derived sample size can ensure the probability of missing important data samples or clusters is low.

Chernoff bound was previously used by Guha et al., in 2001 [15] for reducing the size of large datasets based on sampling. They had used *Chernoff bound* in such a way that the probability of missing samples from a cluster or class is low. It was also proven that sampling-based data summarization reduces the loss of information for a class or a cluster. The bootstrap sampling and *Chernoff bound* theorem are described in the following sections.

3.1. Bootstrap sampling

In 1979, Efron first proposed the bootstrap sampling method in classical statistics for an independent and identically distributed (i.i.d.) sample of fixed size data from a distribution F [20]. After its first introduction, several improvements were proposed for estimating the variance of this method, such as by Wolter [21] and Escobar et al. [22]. Multiple samples are drawn with replacement in the bootstrap sampling method and are usually used for statistical estimation and diagnosis. Formally, the bootstrap sampling method can be described as follows [23]:

Suppose, x_1, x_2, …, x_n is an i.i.d. sample from the unknown distribution f. α is a parameter estimated by $\hat{\alpha}$ - a function of the sample. The bootstrap method then estimates the distribution of $\hat{\alpha}$ calculated from the i.i.d. sample (from the distribution f). The estimation is computed by the distribution of $\hat{\alpha}*$ that is calculated from the i.i.d. bootstrap sample of $x_1^*, x_2^*, \ldots, x_n^*$ from an empirical distribution function $\hat{f}_n$ - an estimate of the f. Hence, the estimate $\hat{f}_n(y)$ for a real number y is defined as

$$\hat{f}_n(y) = \frac{1}{n}\sum\nolimits_{i=1}^{n} I(x_i \leq y) \tag{1}$$

The variance for the bootstrap estimation is $v^* = v^*(\hat{\alpha}^* | x_1, x_2, \ldots, x_n)$ - the conditional variance $\hat{\alpha}^*$ given x_1, x_2, …, x_n. Generally, the bootstrap variance estimator $v*$ is not a closed form function of x_1, x_2, …, x_n. Practically, the Monte Carlo approximation of $v*$ is required to use for this estimation. The bootstrap sampling procedure can be described as follows:

- Generate $x_1^*, x_2^*, \ldots, x_n^* i.\tilde{i}.d.\hat{f}_n$. This is equivalent to a random sample drawn $x_1^*, x_2^*, \ldots, x_n^*$ with a replacement from x_1, x_2, …, x_n.
- If $\hat{\alpha}^*$ is the bootstrap statistic calculated from the resulting bootstrap sample, then repeat the previous step β times (usually a large number) to obtain $\hat{\alpha}_1^*, \hat{\alpha}_2^*, \ldots, \hat{\alpha}_\beta^*$.
- Estimate variance $v(\hat{\alpha})$ as

$$\hat{v}_\beta^* = \frac{1}{\beta - 1}\sum\nolimits_{b=1}^{\beta} \left(\hat{\alpha}_b^* - \hat{\alpha}_{(.)}^*\right)^2 \tag{2}$$

where $\hat{\alpha}_{(.)}^* = \beta^{-1}\sum_{b=1}^{\beta}\hat{\alpha}_b^*$.

When the number of samples β by the bootstrap sampling method goes to infinity based on the original samples, $\hat{v}_\beta^*$ converges close to $v*$ following the law of large numbers.

3.2. Chernoff bound

Consider, a cluster C in a dataset D. Then the probability of summary samples contain a fewer than $f_{CB} \times |C|$ data samples that belong to the cluster C is less than δ ($0 \leq \delta \leq 1$) when the sample size s satisfies equation (3).

$$s \geq f_{CB} \times |D| + \frac{|D|}{|C|} log\left(\frac{1}{\delta}\right) + \frac{|D|}{|C|}\sqrt{\left(log\left(\frac{1}{\delta}\right)\right)^2 + 2 \times f_{CB}|C| log\left(\frac{1}{\delta}\right)} \tag{3}$$

Here, in equation (3), f_{CB} defines the fraction of the cluster $C(0 \leq f_{CB} \leq 1)$.

In this work, the sample size is used to compute the summary size from the original data. It is important to maintain that the ratio of rare patterns present in the original data is also represented in the summary because this summary data is used to identify rare patterns. Accordingly, a modified *Chernoff bound* from Ref. [6] has been used to calculate the summary size for the dataset. The modified *Chernoff bound* theorem is described below:

Modified Chernoff Bound: For an anomalous cluster $C_{anomaly}$, the probability of the summary contains a fewer than $f_{CB} \times |C_{anomaly}|$ data samples that belong to the anomalous cluster is less than $\delta(0 \leq \delta \leq 1)$ when it satisfies equation (4).

$$s_{min} \geq f_{CB} \times |D| + \frac{|D|}{|C_{anomaly}|} log\left(\frac{1}{\delta}\right) + \frac{|D|}{|C_{anomaly}|}\sqrt{\left(log\left(\frac{1}{\delta}\right)\right)^2 + 2 \times f_{CB}|C_{anomaly}| log\left(\frac{1}{\delta}\right)} \tag{4}$$

Based on equation (4), it can be concluded that for the summary data to contain at least $f_{CB} \times |C_{anomaly}|$ samples that belong to the cluster $C_{anomaly}$ with a high probability, the sample needs to contain more than a fraction f_{CB} of the total number of data samples in the dataset. Therefore, if the $C_{anomaly}$ is the smallest cluster, then the $s_{\min}$ is the resulted sample size from equation (4). It can be observed that equation (4) holds for $s = s_{\min}$ and all $|C| \geq |C_{anomaly}|$ [15]. Suppose there is a sample size $s_{\min}$. In that case, it can be guaranteed that the sample or summary data contains at least $f_{CB} \times |C_{anomaly}|$ samples from the anomalous cluster with a high probability $(1 - \delta)$. Hence, based on equation (4), the sample size can be obtained for creating summary data using the bootstrapping sampling.

3.3. Ensemble data summarization Using Chernoff Bound (EDSUCh)

Algorithm 1. EDSUCh: Ensemble Data Summarization Using Chernoff Bound

```
Require: $D$, The dataset; $|C_{anomaly}|$, The size of the anomalous cluster; $\delta$, The probability
    for the sample to contain anomalous samples; $f$, The fraction of the dataset to be
    anomalous cluster; $k$, The value of the parameter $k$ for the $X$-means clustering;
Ensure: $S$, The summary of $D$;
  1: Begin
  2: Apply $X$-means clustering on the dataset to cluster the samples and store in $D_{X-means}$;
  3: Apply $GK$-means clustering on the same dataset to cluster the samples and store in $D_{GK-means}$;
  4:Calculate the sample size $s$ using the modified Chernoff bound (4);
  5: $S_1 \leftarrow$ bootstrapping sample from $D_{X-means}$;
  6: $S_2 \leftarrow$ bootstrapping sample from $D_{GK-means}$;
  7: Combine samples ($S_1$ and $S_2$) together into $S_{combined}$;
  8: Calculate $s_{combined} = \frac{s_{x-means} + s_{GK-means}}{2}$;
  9: Take random samples $S_{random}$ of size $s_{combined}$ from $S_{combined}$;
  10: Remove any duplicate samples from $S_{random}$ and assign it to $S$;
  11: End
```

Algorithm 1 shows the proposed ensemble data summarization using Chernoff bound for effective medical diagnosis. The modified *Chernoff bound* using equation (4) was incorporated with the bootstrapping sampling concept to ensure the rare patterns in the summary. First, X-means and GK-means [24] clustering algorithms were applied to the original datasets separately to cluster the data samples into similar groups. Next, the summary or sample size was calculated using equation (4). Once the sample size is calculated, bootstrapping samples were collected from each clustering (X-means and GK-means) outputs. The samples were then combined, and a random sample was taken. To determine the sample size for the random sampling from combined outputs, half of the total number of samples from each clustering output was considered. Because samples from the original datasets were in each clustering (X-means and GK-means) outputs, after taking the random samples from this combination, duplicate samples were removed. Now, the anomaly (in other words, rare patterns [5] as used in this paper) representations in the summary datasets were identified to see whether the summary data contained any rare patterns or not.

3.4. Proposed methodology for effective medical diagnosis

The proposed methodology for effective medical diagnosis based on the novel EDSUCh is illustrated in Fig. 1. The methodology involves a preprocessing step of converting the collected datasets into a processed format, including CSV and ARFF formatted datasets. Here, multi-class datasets are labelled into two classes - normal and anomalous, where anomalous class contains all the rare patterns except the normal samples. The proposed EDSUCh method is then applied to the processed datasets to create summary datasets following the Algorithm 1. The anomalies present in the summary datasets are identified and compared with the anomaly representability and distribution of anomalous data in original datasets.

The subsequent steps involved applying unsupervised anomaly

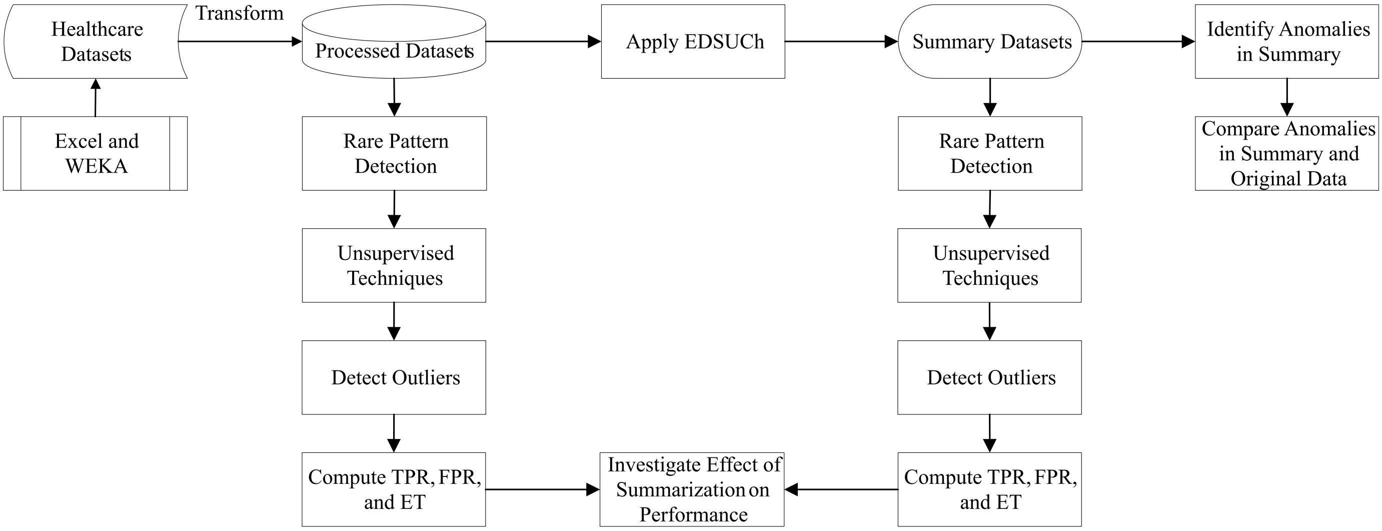


Fig. 1. Proposed methodology using EDSUCh for effective medical diagnosis.

detection algorithms and calculating accuracy (TPR and FPR) both in the summary and original data. The timing (ET) required to summarize and detecting rare patterns were calculated in summary, while only the timing required to detect rare patterns in the original data were calculated. Finally, the accuracy and timing required for rare pattern detection were compared between the summarization and the original dataset. The proposed ensemble approach, EDSUCh, does not require the summary size as input because it already computes the sample sizes using the modified *Chernoff bound*. Thus, it overcomes the limitations of existing summarization techniques. The main challenges in the existing techniques are deciding the summary size and how to represent the original data in the summarized data. The proposed EDSUCh approach represents the original data using a number of data samples instead of the given number of centroids or medoids only. The proposed EDSUCh approach focuses only rare patterns for effective medical diagnosis. It ensures the representability of rare patterns in summary, and hence, it does not pay attention to other patterns.

4. Experimental analysis

The proposed EDSUCh has been evaluated in this section. One of the key aspects of the analysis of the experimental results is to validate the effectiveness of the proposed EDUCh method in representing the rare patterns in the summary and performance improvement in detecting these rare patterns compared to the original datasets for effective medical diagnosis. In experimental analysis, anomaly representability stands for the amount of normal and rare patterns (i.e., anomalous) present in summary.

4.1. Summary of the datasets

A summary of the datasets used in experiments for effective medical diagnosis has been listed in Table 1 with normal, anomalous percentages and number of instances. The four datasets are Breast Cancer Wisconsin (BC), Colon Cancer (CC), Dermatology (DT), and Diabetes (DB). All these datasets contain rare patterns.

4.2. Analysis of results

The proposed method, EDSUCh, identifies the sample size using the modified *Chernoff bound* and it requires three parameters to define. The parameters are $C_{anomaly}$ (a fraction of the dataset contained in the anomalous cluster), δ (a probability to ensure the sample size is less than the fraction of the cluster), and f (a fraction of the anomalous cluster). The values used for these three parameters in experiments are $\delta = 0.90$, $|C_{anomaly}| =$ number of samples in the corresponding cluster, and $f = [0.025, 0.05, 0.1]$. Therefore, three different sets of experiments were conducted for the data summarization.

4.2.1. Representation of rare anomaly in summary

4.2.1.1. Anomaly representability. The proposed ensemble data summarization, EDSUCh, has been evaluated from the perspective of anomaly detection. Table 2 presents the anomaly representability scores for each dataset for three different values of f, which are 0.025, 0.05, 0.1. It can be observed that the percentage of reduction in the selection of anomalous samples, in summary, is much less than the original number of anomalous samples in the original datasets. For the BC dataset, the percentages of reduction are 92.92%, 84.91%, and 70.28%, respectively. For the CC dataset, the percentages are 82.50%, 82.50%, and 67.50%, respectively. For the DT dataset, these are 85.83%, 77.17%, and 62.99%, respectively. Finally, for the DB dataset, the percentages are 95.00%, 90.20%, and 80.60%, respectively as shown in Table 2.

4.2.1.2. Distribution of anomalies in original and summary data. To further discover the effectiveness of the proposed ensemble summarization method, EDSUCh, the distribution of rare patterns in the original data and in summary is presented in Table 3. The distribution of rare patterns in the original data (D_R) is calculated as in equation (5) where N_R is the number of rare patterns in original dataset and N_D is the number of data instances in the original dataset.

Table 1
Distribution of normal and anomalous data.

Dataset	Normal (%)	Anomaly (%)	No. of instances
BC	62.74	37.26	569
CC	35.48	64.52	62
DT	30.60	69.40	366
DB	34.90	65.10	768

Table 2
Anomaly representability comparison.

Dataset	**Original**	**f=0.025 (%)**	**f=0.50**	**f=0.1**
BC	212	15	32	63
CC	40	7	7	13
DT	254	36	58	94
DB	500	25	49	97

Table 3
Distribution of anomaly (%) in original data and summaries.

Dataset	Original (%)	f=0.025 (%)	f=0.50 (%)	f=0.1 (%)
BC	37.26	45.45	50.79	55.26
CC	64.52	77.78	77.78	76.47
DT	69.40	83.72	76.32	77.69
DB	65.10	58.14	59.76	62.18

$$D_R = \frac{N_R}{N_D} \tag{5}$$

and the distribution of rare patterns in summary (S_R), is calculated as in equation (6) where N_{RS} is the number of rare patterns in summary and N_S is the number of data instances in the summary.

$$S_R = \frac{N_{RS}}{N_S} \tag{6}$$

It can be observed from Table 3 that the data distribution of the rare patterns, in summary, is higher in all cases for all datasets. For the BC dataset, the increased percentages of rare patterns in summary compared to the original data are 21.98%, 36.31%, and 48.31%, respectively. For the CC dataset, these are 20.55%, 20.55%, and 18.52%, respectively. For the DT dataset, the percentages are 20.63%, 9.97%, and 11.95%, respectively. Finally, for the DB dataset, these are 10.69%, 8.20%, and 4.49%, respectively.

4.2.2. Effectiveness of proposed ensemble data summarization method for effective medical diagnosis

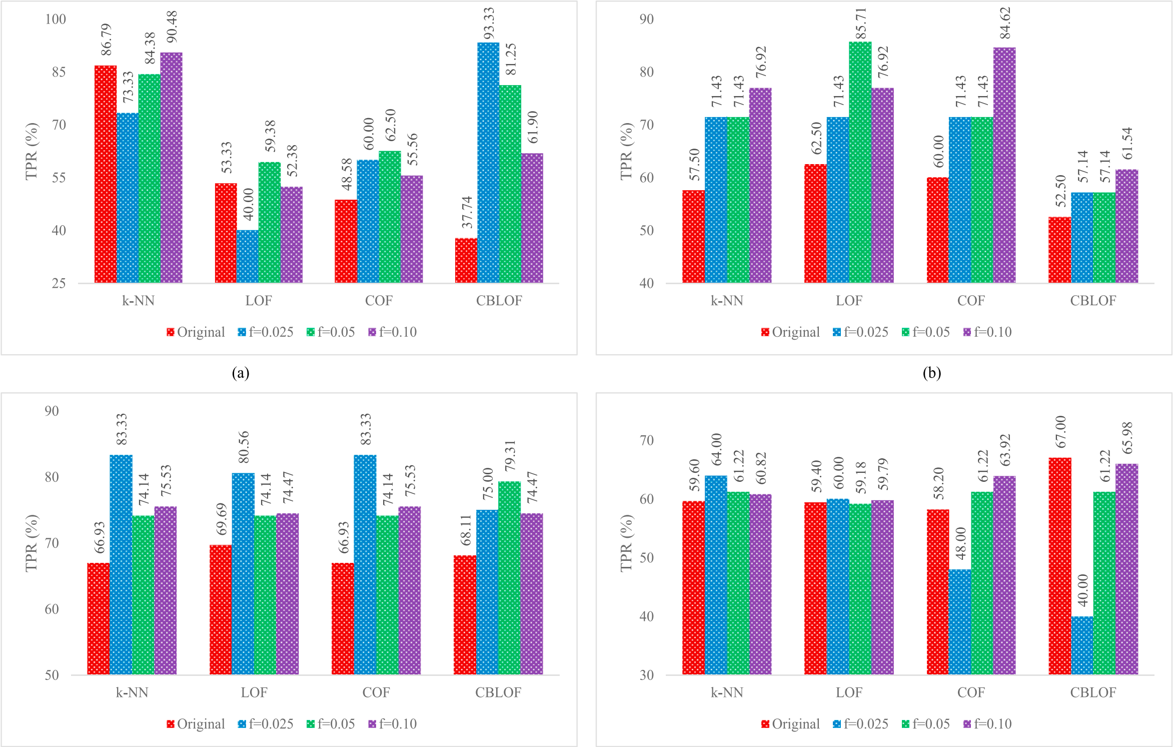
Anomaly detection methods are usually used to handle different types of anomalous activities, such as attacks in network traffic analysis. Similar techniques are also used to detect rare patterns in medical diagnosis. Three dominant approaches, including supervised, unsupervised, and semi-supervised, are widely used in different domains for anomaly detection [25,26]. Because data summarization is used as a preprocessing step before the anomaly detection techniques are applied to improve the computational efficiency and to improve the performance, in this paper, unsupervised anomaly detection techniques are explored rather than the supervised and semi-supervised methods. Interested readers can study the popular unsupervised anomaly detection techniques in Refs. [27,28].

Fig. 2 shows the performance in terms of TPR for all individual anomaly detection techniques, *k*-NN, LOF, COF, and CBLOF, using original data and summaries for all datasets. It can be observed that with an exception for the Diabetes dataset in the case of CBLOF, the performance of the summary is improved over the original dataset.

Fig. 3 (a) shows the average TPR performance for the *k*-NN, LOF, COF, and CBLOF anomaly detection techniques, using original data and summaries for all datasets. It can be observed that all anomaly detection techniques' performance was improved when using the ensemble data summarization approach, EDSUCh, and for any value of the parameter *f* when calculating the sample size. Fig. 3 (b) shows the average TPR performance for all datasets using original data and three different summaries. It can be observed that all summaries performed better than the original data for detecting rare patterns for effective medical diagnosis. It can be noted that among the three different summaries, the highest overall performance was achieved by the summary with sample size $f = 0.05$. However, it was the performance was very close to the other summary with sample size calculated using $f = 0.10$. It can be concluded that if the summary contained a higher fraction of the anomalous cluster, the performance of anomaly or rare pattern detection could be improved at a substantial amount.

4.2.3. Comparison of computational time

While data summarization can improve the anomaly detection performance compared to the original data, the combined computational time required to summarize the data and anomaly detection should also be less than the anomaly detection using the original dataset. Fig. 4 shows the timing performance of the combined timing required for



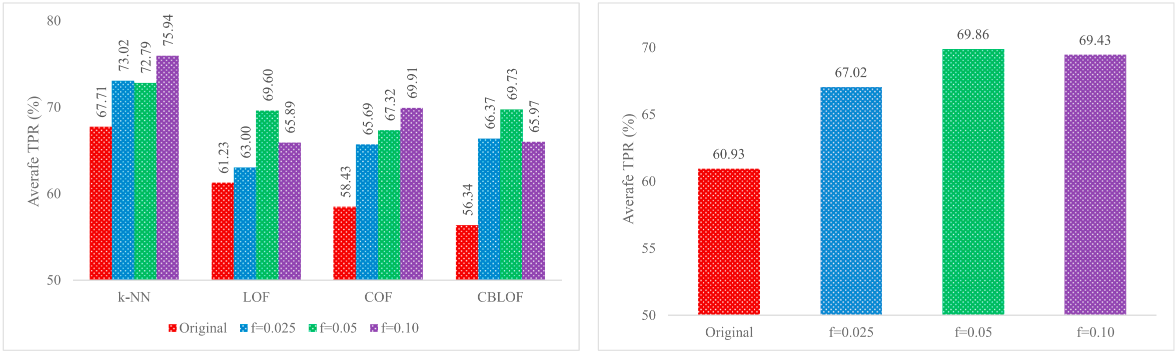
(a)

(b)

(c)

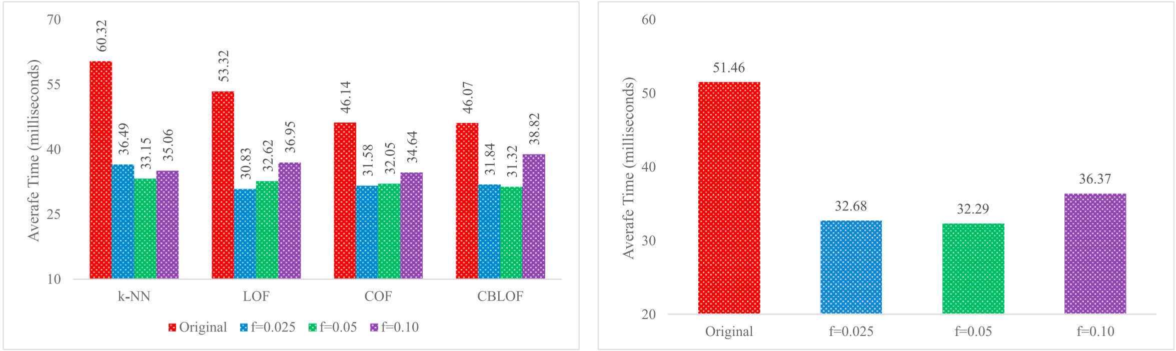
(d)

Fig. 2. Performance of the individual anomaly detection techniques (*k*-NN, LOF, COF, and CBLOF) using original data and summary for all datasets. (a) Breast Cancer Wisconsin (b) Colon Cancer (c) Dermatology (d) Diabetes.



(a) (b)

Fig. 3. (a) Average performance of the individual anomaly detection techniques (*k*-NN, LOF, COF, and CBLOF) using original data and summary for all datasets. (b) Average performance of all datasets using original data and summary.



(a) (b)

Fig. 4. (a) Average timing performance of the individual anomaly detection techniques (*k*-NN, LOF, COF, and CBLOF) using original data and summary for all datasets. (b) Average timing performance of all datasets using original data and summary.

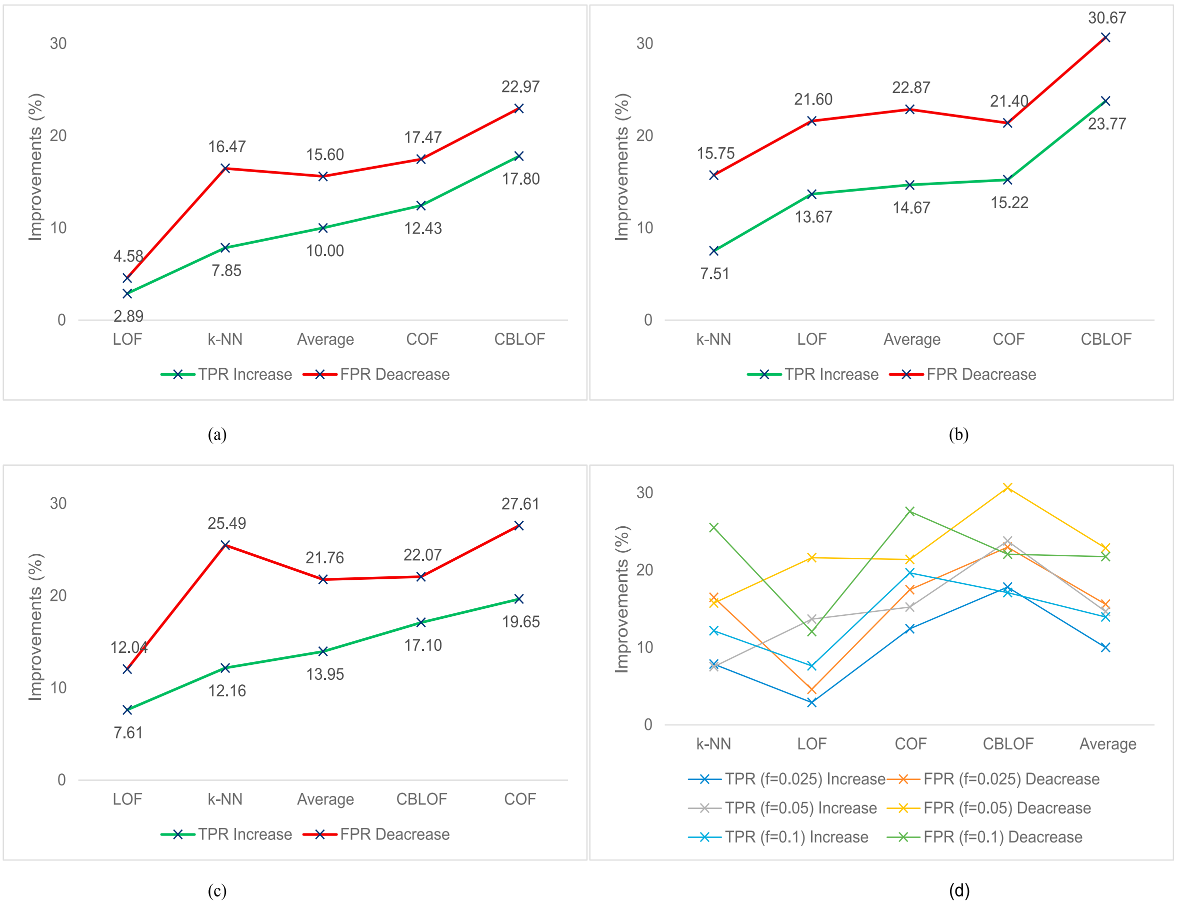
summarizing the data and anomaly detection versus the only anomaly detection using original data. From Fig. 4 (a), it can be observed that the combined time required for creating the summary and for detecting rare patterns is always less than the only time required for detecting rare anomalies using the original datasets for all the underlying anomaly detection algorithms. Fig. 4 (b) shows the average timing required for all datasets by all anomaly detection techniques for the original data and the EDSUCh method. Likewise, the improved TPR performance by EDSUCh when the sample size was calculated using $f = 0.05$, the average timing required by EDSUCh for all datasets using the summary was less than others when the sample size was calculated for $f = 0.50$.

4.2.4. Key insights

Fig. 5 shows the key insights by the proposed ensemble data summarization using modified *Chernoff bound*, EDSUCh, for effective medical analysis. From the experimental results, it was shown that the underlying all anomaly detection techniques' (*k*-NN, LOF, COF, and CBLOF) performance improved when using the EDSUCh method for detecting the rare anomalies from the medical datasets (Breast Cancer Wisconsin, Colon Cancer, Dermatology, and Diabetes). Fig. 5 (a) shows the performance improvements of the anomaly detection techniques in terms of TPR increase and FPR decrease when the data sampling has been calculated using $f = 0.025$. Fig. 5 (b) shows the TPR increase and FPR decrease performance improvements for $f = 0.05$ and Fig. 5 (c) shows the same performance improvements for $f = 0.1$. It can be observed that the performance of the different algorithms was the highest compared to others when the data sampling was performed using the Bootstrapping sampling with modified *Chernoff bound*. For example, when $f = 0.025$, the highest average TPR was increased by CBLOF, while the lowest average TPR was achieved by LOF. When $f = 0.05$, the highest and lowest TPR performance was improved by CBLOF and *k*-NN, respectively. Finally, when $f = 0.1$, the highest and lowest TPR was obtained by COF and LOF algorithms, respectively. Likewise, the same algorithms achieved the TPR performance improvement, the highest and lowest FPR performances. Fig. 5 shows the combined TPR increase and FPR decrease performances for all values of f. It can be seen that while the lowest achievable TPR increase and FPR decrease performances were by LOF, the highest TPR increase and FPR decrease performances were achieved by the CBLOF anomaly detection technique in most cases of different values of f used for calculating the sample size for data summarization. Therefore, it can be observed that the performance of all underlying algorithms was improved substantially when using the proposed ensemble data summarization method, EDSUCh, for summarizing the data before analyzing the data for rare pattern detection for effective medical analysis.

5. Discussion and conclusion

In this paper, an ensemble data summarization method was proposed for effective medical diagnosis. The aim of this ensemble approach was to reduce the complexity of medical diagnosis on original data, such as the detection of rare patterns. The proposed ensemble data summarization method, EDSUCh, was based on data sampling. In the process of data summarization, the crucial part was how to find the best summary. While the optimal size of a summary impacts the quality of the summary, it is also important to represent the underlying data patterns. Therefore, an appropriate data summarization method can be an obvious solution for effective medical diagnoses, such as anomaly detection or the detection of rare patterns. Throughout the experiments in this paper, it has been shown that using the proposed EDSUCh method, the performance of rare pattern detection yielded better when the summary data is used



(a)

(b)

(c)

(d)

Fig. 5. Key insights of ensemble data summarization (EDSUCh) for effective medical diagnosis via performance improvements in terms of TPR increase and FPR decrease percentages. (a) $f = 0.025$. (b) $f = 0.05$. (c) $f = 0.1$. (d) Combined.

compared to the original data.

The objective of data summarization techniques is to produce a concise summary of the original data so that it can represent the original data and be used for later data analysis tasks, such as rare pattern detection in network traffic analysis or effective medical diagnosis. Data summarization helps to make the anomaly detection techniques more scalable and efficient compared to using original data. This eventually benefits the data analysts for making effective decisions to optimize their underlying system. In this paper, ensemble data summarization and rare pattern detection for effective medical analysis have been considered. Based on the experimental results and analysis presented in this paper, it can be adapted to any domain, including cybersecurity and financial domains, such as anomaly detection in cybersecurity, with a reduced execution time and improved detection performance.

An important aspect of the data summarization process is to preserve the anomaly distribution in summary compared to the original dataset. A balanced anomaly distribution, in summary, can produce better performance results, as observed throughout the experimental analysis in this paper. However, in summary, preserving anomalous samples can be more challenging when the original dataset contains collective anomalies. Collective anomalies act as a group, and the normal samples can be of the nature of anomalous samples. Therefore, the proposed ensemble data summarization method, EDSUCh, can be studied for collective anomalies in future work. Furthermore, the proposed approach can be experimented with in other domains, such as cybersecurity and financial domains, to prove its effectiveness without a boundary. The ensemble data summarization's effectiveness may vary in computations and anomaly detections for different domains, which should be experimented with that domain's data and remain a limitation of this current study.

Declaration of competing interest

The authors declare that there is no conflict of interest.

References

[1] A.N.M.B. Rashid, M. Ahmed, L.F. Sikos, P. Haskell-Dowland, A novel penalty-based wrapper objective function for feature selection in Big Data using cooperative co-evolution, IEEE Access 8 (2020) 150113–150129.

[2] A.N.M.B. Rashid, M. Ahmed, L.F. Sikos, P. Haskell-Dowland, Cooperative co-evolution for feature selection in big data with random feature grouping, J. Big Data 7 (1) (2020) 1–42.

[3] A.N.M.B. Rashid, M. Ahmed, L.F. Sikos, P. Haskell-Dowland, Anomaly detection in cybersecurity datasets via cooperative co-evolution-based feature selection, ACM Transact. Manag. Inform. Syst. 13 (3) (2022) 1–39.

[4] A.N.M. Rashid, M. Ahmed, S.R. Islam, A supervised rare anomaly detection technique via cooperative co-evolution-based feature selection using benchmark unsw_nb15 dataset, in: International Conference on Ubiquitous Security, Springer, 2021, pp. 279–291, https://doi.org/10.1007/978-981-19-0468-4_21.

[5] A.N.M.B. Rashid, M. Ahmed, A.-S. K. Pathan, Infrequent pattern detection for reliable network traffic analysis using robust evolutionary computation, Sensors 21 (9) (2021) 3005.

[6] M. Ahmed, Intelligent big data summarization for rare anomaly detection, IEEE Access 7 (2019) 68669–68677.

[7] P. Lavanya, K. Kouser, M. Suresha, Effective feature representation using symbolic approach for classification and clustering of big data, Expert Syst. Appl. 173 (2021) 114658.

[8] A. Bharadwaj, A. Srinivasan, A. Kasi, B. Das, Extending the performance of extractive text summarization by ensemble techniques, in: 2019 11th International Conference on Advanced Computing, ICoAC, 2019, pp. 282–288, https://doi.org/10.1109/ICoAC48765.2019.246854.

[9] A. Onan, Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering, IEEE Access 7 (2019) 145614–145633.

[10] T. Boongoen, N. Iam-On, Cluster ensembles: a survey of approaches with recent extensions and applications, Computer Sci. Rev. 28 (2018) 1–25.

[11] Z. Liao, L. Gao, T. Zhou, X. Fan, Y. Zhang, J. Wu, An oil painters recognition method based on cluster multiple kernel learning algorithm, IEEE Access 7 (2019) 26842–26854.

[12] J. Wu, S. Guo, H. Huang, W. Liu, Y. Xiang, Information and communications technologies for sustainable development goals: state-of-the-art, needs and perspectives, IEEE Commun. Surveys Tutorials 20 (3) (2018) 2389–2406.

[13] J. Wu, S. Guo, J. Li, D. Zeng, Big data meet green challenges: big data toward green applications, IEEE Syst. J. 10 (3) (2016) 888–900.

[14] R. Atat, L. Liu, J. Wu, G. Li, C. Ye, Y. Yang, Big data meet cyber-physical systems: a panoramic survey, IEEE Access 6 (2018) 73603–73636.

[15] S. Guha, R. Rastogi, K. Shim, Cure: an efficient clustering algorithm for large databases, Inf. Syst. 26 (1) (2001) 35–58.

[16] A.N. Mahmood, J. Hu, Z. Tari, C. Leckie, Critical infrastructure protection: resource efficient sampling to improve detection of less frequent patterns in network traffic, J. Netw. Comput. Appl. 33 (4) (2010) 491–502.

[17] M. Ahmed, A.N. Mahmood, M.J. Maher, An efficient technique for network traffic summarization using multiview clustering and statistical sampling, EAI Endorsed Transact. Scalable Informat. Syst. 2 (5) (2015) e4.

[18] M.S. Mahmud, J.Z. Huang, S. Salloum, T.Z. Emara, K. Sadatdiynov, A survey of data partitioning and sampling methods to support big data analysis, Big Data Min. Analy. 3 (2) (2020) 85–101.

[19] D. Gillman, A chernoff bound for random walks on expander graphs, SIAM J. Comput. 27 (4) (1998) 1203–1218.

[20] B. Efron, Bootstrap methods: another look at the jackknife, in: Breakthroughs in Statistics, Springer, 1992, pp. 569–593.

[21] K.M. Wolter, K.M. Wolter, Introduction to Variance Estimation, vol. 53, Springer, 2007, https://doi.org/10.1007/978-0-387-35099-8.

[22] E.L. Escobar, Y.G. Berger, A jackknife variance estimator for self-weighted two-stage samples, Stat. Sin. (2013) 595–613.

[23] Z. Mashreghi, D. Haziza, C. Léger, A survey of bootstrap methods in finite population sampling, Stat. Surv. 10 (2016) 1–52.

[24] M.Z. Islam, V. Estivill-Castro, M.A. Rahman, T. Bossomaier, Combining k-means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering, Expert Syst. Appl. 91 (2018) 402–417.

[25] M. Ahmed, A.N. Mahmood, J. Hu, A survey of network anomaly detection techniques, J. Netw. Comput. Appl. 60 (2016) 19–31.

[26] M. Ahmed, A.N. Mahmood, M.R. Islam, A survey of anomaly detection techniques in financial domain, Future Generat. Comput. Syst. 55 (2016) 278–288.

[27] M. Ahmed, A. Anwar, A.N. Mahmood, Z. Shah, M.J. Maher, An investigation of performance analysis of anomaly detection techniques for big data in scada systems, EAI Endorsed Transact. Industr. Networks Intelligent Syst. 2 (3) (2015) e5.

[28] M. Ahmed, Reservoir-based network traffic stream summarization for anomaly detection, Pattern Anal. Appl. 21 (2) (2018) 579–599.