Journal of Big Data

RESEARCH Open Access

Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest

Maria Irmina Prasetiyowati*, Nur Ulfa Maulidevi and Kridanto Surendro

*Correspondence:
maria@umn.ac.id
School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia

Abstract

Feature selection is a pre-processing technique used to remove unnecessary characteristics, and speed up the algorithm's work process. A part of the technique is carried out by calculating the information gain value of each dataset characteristic. Also, the determined threshold rate from the information gain value is used in feature selection. However, the threshold value is used freely or through a rate of 0.05. Therefore this study proposed the threshold rate determination using the information gain value's standard deviation generated by each feature in the dataset. The threshold value determination was tested on 10 original datasets transformed by FFT and IFFT and classified using Random Forest. On processing the transformed dataset with the proposed threshold this study resulted in lower accuracy and longer execution time compared to the same process with Correlation-Base Feature Selection (CBF) and a standard 0.05 threshold method. Similarly, the required accuracy value is lower when using transformed features. The study showed that by processing the original dataset with a standard deviation threshold resulted in better feature selection accuracy of Random Forest classification. Furthermore, by using the transformed feature with the proposed threshold excluding the imaginary numbers leads to a faster average time than the three methods compared.

Keywords: Threshold, Standard deviation, Accuracy, Time, Random forest

Introduction

Data development increases dimensions and computational costs are overcome by feature selection and extraction, which are two different techniques [1, 2]. Several studies have shown the ability of the feature extraction process to build a new feature set [3–6]. Conversely, the selection involves choosing a subset of the original feature set [1]. It is also a technique used in pre-processing, which affects model performance by removing excessive and unimportant features [7]. Moreover, the technique speeds up the work process of an algorithm [8]. There are several means of performing feature selection [9], broadly grouped into three main types: Filter [10–13] Wrapper [14, 15], and Embedded [16]. The filter method evaluates feature subsets with predefined criteria independent of any grouping [17]. Information Gain (IG) is a popular filter model and technique used

in feature weight scoring and to determine the maximum entropy value. However, as a basic technique, IG is still open to further research and development in feature selection. Elmaizi [18] proposed a new approach based on IG for image classification and dimension. Similarly, Jadhav [19] proposed feature selection based on IG ranking based, while Singer [20] developed a model known as Weighted Information-Gain (WIGR), which defines proportionally weighted entropy.

All dataset features in IG were counted, selected, and defined by a value limit known as the threshold (cutoff). The threshold value of 0.05 [21, 22] is often set freely as required and used whenever a study requires good accuracy at a lower level. Tsai and Sung researched by calculating the average of each frequency to obtain the threshold value of the final features' subset [23]. Preliminary studies determined the acquired threshold value according to the standard deviation of the IG rate. Furthermore, each feature's weighting result was calculated, while the threshold value was determined using the standard deviation. However, in this study, the standard deviation is used to express the diversity of IG value distribution with Information Gain chosen because of its ability to measure the data possessed by each feature. The IG method is used in decision trees to maximize the richness of information. This study used a simpler method with original and transformed data sets, while the transformation of each feature's IG value was performed using the FFT to accelerate the algorithm's performance. In addition, this study comprises ten datasets with more than 100 features (high-dimensional datasets) compared to others with less features. Several studies do not consider the speed of execution in Random Forest usage.

The Random Forest is a tree-based learning algorithm machine, which leverages the power of multiple decision trees for making decisions [24]. The feature selection in Random Forest calculations is selected more than once, and this involves a haphazard process that requires a very long computational time. Moreover, the feature selected to construct a decision tree may not be informative.

Fast Fourier Transform (FFT) is an algorithm applied to increase execution speed. This recursive method involves dividing the original vector into two parts, combining and calculating their individual FFT. Several studies stated that FFT enhances execution and is used in feature extraction methods. For instance, Herf [25] stated that FFT could be used on a dataset of time series collected in sequential time series, such as clinical data [26, 27]. The application of the FFT algorithm to the dataset will not change the data, because the IFFT algorithm will returns the dataset to its original data. Prasetiyowati et al. [8], in his research analyzed this method and produced the accuracy value and time needed better than the original dataset. Therefore, based on the results of these studies, FFT is applied in this study. Therefore, based on the results of these studies, FFT is applied in this study. Besides being applied to the dataset, FFT is also applied to feature extraction [28, 29], and selection [30]. Gowid et al. [30] used used this process to develop robust, fast, and automated feature selection algorithms for mechanical systems. Based on these studies, the FFT algorithm is also used to perform feature selection. Data and Information Gain values for each feature are transformed as a signal wave with various values. Differ from the previous research is that this study examines whether the dataset and features transformed by FFT and IFFT produce better accuracy & average speed values compared to choosing the Correlation-Base Feature Selection model and threshold

of 0.05. Also, FFT and IFFT were generally used on image or signal datasets, while in this studies both are used for non-image information.

This study follows previous research on the use of feature selection to increase the Random Forest method performance on high dimensions [31]. It also examines the speed and accuracy evaluation of Random Forest performance by selecting features in the transformation data [8].

The key contribution of this paper is provided as follows,

- Propose a feature selection method in Random Forest.
- The proposed feature selection method is Information Gain, using a threshold with a standard deviation calculation,
- Compares the mean value of Random Forest accuracy and speed from the results, with standard deviation, Correlation-Base Feature Selection, and threshold of 0.05,
- Compares the mean value of Random Forest accuracy and speed from the results, using the original and transformed dataset, through FFT and IFFT,
- Compares the mean Random Forest speed and accuracy, using features transformed with FFT.

The research is divided into several sections. The 2nd, 3rd, 4th, 5th, and 6th sections are the related work, the proposed method, the research results, discussion, and conclusion, respectively.

Related work

Information gain

Information Gain (IG) is an entropy-based selection method [32], which involves the calculation from the output data grouped by feature A, denoted as gain (y, A). The Information Gain (y, A) is represented as,

$$gain(y, A) = entropy(y) - \Sigma_{C \in vals(A)} \frac{y_c}{y} entropy(yc) \quad (1)$$

The value (A) is the possible rates of attribute A, with Yc being the subset of y, where A possesses the sum of c. Furthermore, the rule of Eq. (1) was the total entropy of y, followed by data segregation, based on feature A.

Studies are still carried out on the development of Information Gain to date. An instance is a study conducted by Elmaizi [18], which proposed a new approach based on IG for hyperspectral image classification and dimensional reduction. The hyperspectral band selection was used to select the most informative ribbons and remove irrelevant and noisy bands. The comparison results showed that the information retrieval filter approach was superior, reduced computational costs, and enhanced classification accuracy. Moreover, the dataset used were two of the hyperspectral images obtained from The Indian Pines AVIRIS and The Pavia University. This study is in contrast with the research carried out by Jadhav et al. [19], which proposed feature selection through ranking based on IG. Furthermore, the technique used was known as the Information Gain Directed Feature Selection algorithm (IGDFS), which is a method that makes use of feature ranking, based on data acquisition through the GA wrapper (GAW), and three

classic KNN machine learning algorithms, Naïve Bayes, and Support Vector Machine (SVM). Furthermore, this method reduces computation costs and improves classification accuracy. It only uses 3 datasets with less than 100 features, including The German (20 features), The Australian (14 features), and Taiwan (24 attributes) credit datasets. Singer [20] proposed a model that defined proportionally weighted entropy known as Weighted Information-Gain (WIGR). The method used was measured through a weighted entropy function that was defined proportionally with different target class values. Singer's study used 12 datasets with less than 100 features (min. 7 and max. 32).

Threshold

Threshold, also known as a threshold (cutoff), is the value used as a reference for the selected feature in IG. The threshold value is determined independently or uses a value of 0.05. Tsai and Sung used calculations and averaged each frequency to obtain a final feature's threshold value [23]. Tsai's idea allows the determination of the threshold value using standard deviation.

The process of determining the data group diversity involves reducing the information value through the association's mean and adding the results. This method is known as standard deviation and describes the difference between the measured data against the average value. In this study, the data group is the IG value for each feature in a dataset, which is obtained using Eq. (2).

$$S = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_1)^2}{n(n-1)}} \tag{2}$$

where S is the standard deviation, x is the average value of the IG, xi is the rate of x to i, and n is the number of features used in the dataset.

Random forest

Random Forest is an extension of the decision tree approach developed by Breiman [33–35] and Cutter. It is a tree-based machine learning algorithm that harnesses the power of multiple decision trees in decision-making. The feature selection in the Random Forest calculation is performed carelessly and more than once and requires a very long computation time. As a result, this random process allows the selected feature to be non-informative.

Several preliminary studies examined feature selection for Random Forest. For instance, Yunming's [28] study proposed a stratified sampling method to select feature subspace for Random Forest with high dimensional data as well as strong and weak informative features.

Fast fourier transform and inverse fast fourier transform

An effective medium of converting time-domain signals to frequency domination is the Fast Fourier Transform (FFT) [36] with the Inverse Fast Fourier Transform (IFFT) algorithm used to convert data to the original domain.. Furthermore, the test of the transformed data using FFT and IFFT is applied to high-dimensional and regular datasets with less than 100 features. Hamid used the FFT algorithm to enhance the classification results

through extraction and signal processing [37]. Additionally, Herff [25] stated that clinical data is the time series information often processed and collected sequentially and presented in a continuous waveform using the FFT algorithm. Prasetiyowati et al. [8] also researched the application of the FFT using the Correlation Base Feature Selection. The result showed that the transformed dataset produces an average accuracy and time value better than the original. The transformed dataset is returned using IFFT. In addition, other studies used the FFT algorithm to perform feature extraction [28, 29, 37]. For instance, Ansari used Fast Fourier Transform (FFT) to extract features in the EEG dataset [28], with better accuracy than other classifiers. Meanwhile, in another study, Gowid et al. used FFT for feature selection in mechanical systems [30], with a detection accuracy of 100%.

The FFT and IFFT are shown in Eqs. 3, 4, and 5, respectively.

$$X[k] = \sum_{n=0}^{N-1} X[n] W_N^{kn},\ \mathrm{k} = 0,\ 1,\ \ldots \mathrm{N} - 1, \tag{3}$$

When X[k and n] is a complex number, and W_N^{kn} is the Twiddle factor. N is the order and kn is the index. Defined as follows

$$W_N^{kn} = e^{-j\frac{2xkn}{N}} = \mathrm{Cos}\left(\frac{2\pi kn}{N}\right) - j\sin\left(\frac{2\pi kn}{N}\right) \tag{4}$$

j is an imaginary number, index n is the time variable t in discrete form, and k is the frequency transformation pair

Inverse Fast Fourier Transform (IFFT) is the opposite of FFT, a fast algorithm for calculating IDFT (Inverse Discrete Fourier Transform). IFFT is also calculated using the direct FFT algorithm and complex conjugates

$$\mathrm{X[n]} = \frac{1}{\mathrm{N}} \sum_{\mathrm{k}=0}^{\mathrm{N}-1} \mathrm{X[k]W_N^{-kn}},\ \mathrm{n} = 0, 1, \ldots \mathrm{N} - 1 \tag{5}$$

where X [n] is the inverse of X [k], using the opposite sign and multiplied by a factor of 1/N.

The proposal methods

The search for features in the Random Forest allows the selected feature to be formative. Therefore, a feature search method is needed before executing Random Forest to ensure the features are informative, speed up execution, and increases accuracy.

Information Gain is the proposed feature search obtained using a threshold based on the standard deviation value using Eq. 1.

$$gain(y, A) = entropy(y) - \Sigma_{C \in vals(A)} \frac{y_c}{y} entropy(yc)$$

After determining the Information Gain value for each feature, the next step is to obtain the standard deviation value from the data using Eq. 2.

$$S = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_1)^2}{n(n-1)}}$$

When x is the average value of the Information Gain (IG), and n is the number of features used in the dataset. When searching for the standard deviation value, which also acts as the threshold, the Mean value of the Information Gain automatically appears. This is because before looking for the standard deviation, the mean is first determined. The Information Gain value for each feature is calculated, and the standard deviation is determined as the threshold limit. Furthermore, an IG value equal to or above the threshold value is selected as an informative feature for later use in the Random Forest. The overall framework of the proposed method is shown in Fig. 1.

Workflow

This research consisted of 3 stages. In the first, the experiment uses original data. Second, change the dataset by using FFT and IFFT. And third, change the features using FFT. Next, compare the results of these steps by performing a feature transformation using FFT. These steps are shown in the pseudo-code of the research stage algorithm, including the process of calculating speed and accuracy.

Data were obtained by collecting 10 existing datasets from the UCI and Kaggle, with 3 used in previous studies [8]. The ten datasets were checked for missing values, and in the presence of any, data is completed by giving a zero value [38–40]. The next step involved checking whether the dataset needed transformation with the existing dataset transformed using the FFT algorithm and returned using IFFT. Furthermore, the time and accuracy required to execute each dataset were calculated. However, when the dataset needed no transformation, Random Forest prediction's time and accuracy value was immediately calculated. The pseudo-code of the research stages algorithm, speed, and accuracy calculations were listed in Algorithm 1 and 2.

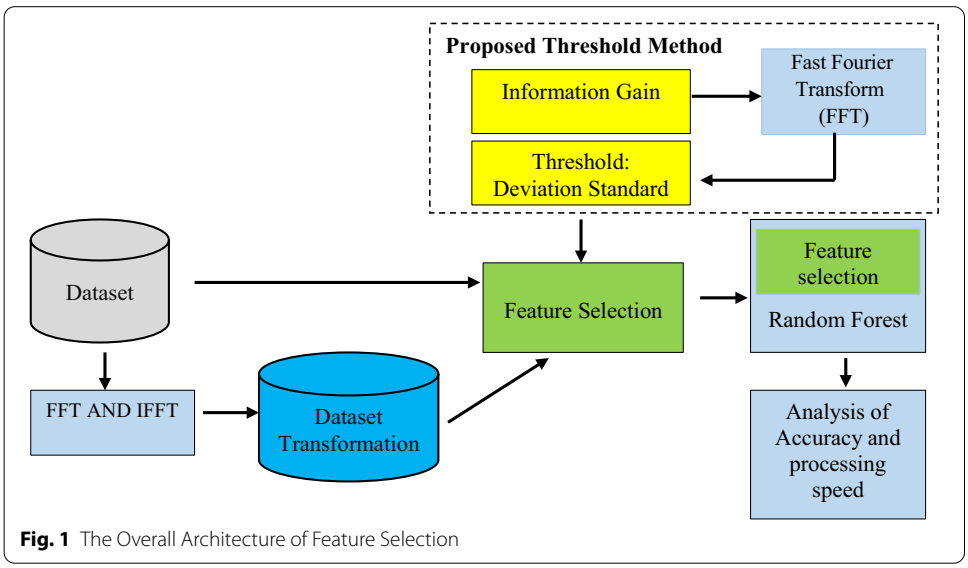

Fig. 1 The Overall Architecture of Feature Selection

```
Pseudocode Research_Stages
Begin
  Input Dataset
  IF Dataset = Trasnformation
      FFT_And_IFFT
  Calculate_Speed_And_Accuracy
  Compare Speed FFT Dataset And Speed Original Dataset
  Compare Accuration FFT Dataset And Speed Original Dataset
End
```

Algorithm 1. Research stages algorithm.

```
Pseudocode Calculate_Speed_And_Accuracy
Begin
     Input Dataset
    Calculate Information Gain
    Write Information Gain
    Calculate Standard Deviation
    Threshold = Standard Deviation Value
    IF Information Gain >=Threshold
        Feature selected = Feature
        For i=1 To I <=10
           Input seed
           Run Predict Random Forest
           Calculate Speed
           Calculate Accuracy
           Write Speed, Accuracy
        EndFor
 ELSE
      Delete Feature
EndIF
End
```

Algorithm 2. The proposed algorithm and calculate speed and accuracy

The calculation of the needed accuracy and time value requires the following steps,

1. Collection of dataset,
2. Selection of the IG feature through the Ranker method, by using the Weka machine learning tools (version 3.9.2).

$$Entropy(y) = -sum(Pi * \log 2(Pi))$$

$$gain(y, A) = entropy(y) - \Sigma_{C \in vals(A)} \frac{y_c}{y} entropy(yc)$$

3. Calculating the standard deviation of the IG for each feature.
   The value obtained during the calculation of all features' standard deviation is known as the threshold. All features possessing a value greater than or equal to that of the threshold value should be selected. Those having a lesser value than that of the threshold should be discarded.

$$S = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_1)^2}{n(n-1)}}$$

4. Performing the selection process, by removing features having an IG value below that of the threshold.
5. The results of the feature selection were used to carry out the Random Forest prediction process, through the Cross-Validation Method Fold = 10.

The Random Forest prediction process was carried out, using 10 randomly selected seed values, including 1, 33, 57, 70, 153, 251, 300, 457, 505, and 700. Therefore, the aforementioned steps were conducted for each dataset.

After the selection process using the Random Forest algorithm, the test results with the proposed threshold were compared with the prediction outcomes using Correlation-Base Feature Selection (CBFS) and the threshold 0.05 technique. In addition, the comparisons were performed using the original and transformed datasets.

Experiment result

Data experiment

The experimental environment is a computer with an Intel processor of 1.60 GHz, 1800 MHz, 4 Core (s), 8 Logical Processor with 12 GB RAM, and 1000 Gigabyte hard drive capacity. This study used a multivariate dataset, real, categorical integer characteristics, and the text numeric converted values. Almost all datasets have complete data except the Dermatology, which had missing values. It is necessary to pre-process the dataset by entering a zero value in the missing data. This study selected some Life, Physical, and Business datasets in the UCI Machine Learning Repository [41]. The Life area dataset consists of EEG Eye, Cancer [42], Contraceptive Method, Dermatology, Divorce [43], Epilepsy [44], and SCADI [45, 46]. While the Physical area dataset consists of the Electrical Grid and Urban Land Cover dataset [47, 48]. The dataset for the business area is the CNAE-9.

This research consisted of 3 stages. In the first, the experiment uses original data. Second, change the dataset by using FFT and IFFT. And third, change the features using FFT. Next, compare the results of these steps by performing a feature transformation using FFT.

The third process produces complex numbers (real and imaginary). Furthermore, to determine the threshold, the authors used two methods, namely:

1. Calculating the standard deviation using real and imaginary numbers.
2. Calculating the standard deviation using only real numbers (without imaginary).

Result

This study used the Random Forest method and K-Cross Fold Validation, with a value of K = 10. Each dataset used ten randomly generated seed values, namely 1, 33, 57, 70, 153, 251, 300, 457, 505, and 700. Furthermore, tests were carried out on the original and transformed dataset through the use of FFT and IFFT methods. Before

transforming datasets with FFT, always ensure that the information for each feature is numeric, with no missing values. This study used the IG as the feature selection technique, with the value of each feature used to calculate the standard deviation parameter as a threshold determination. Furthermore, feature selection was applied to the datasets used, which were then analyzed by utilizing the Random Forest technique. Then, compare the average value of accuracy and time required by the three methods, namely the standard deviation threshold method, the CBFS method, and the 0.05 Threshold method. This comparison also applies to features that have been transformed and use the threshold method based on standard deviation.

Accuracy and speed dataset

EEG dataset The first test was conducted on the EEG Eye dataset, which had 14,980 instances with 14 features. Through IG using the proposed threshold, the standard deviation value is 0.0171, the average value is 0.0289, and 10 features are selected. Furthermore, the resulting average accuracy value was higher (90.15%), and outperforming the rate generated by the CBFS and threshold 0.05. However, the average execution time was faster, through the use of the Correlation-Base Feature Selection. Also, the distinction of the average time was 0.74 secs.

Meanwhile, on the test using a transformed dataset and proposed threshold, the resulting standard deviation value was 0.0171, with 10 selected features. The resulting average accuracy value was higher (90.14%), and also outperforming the rate generated by the CBFS and threshold value of 0.05. However, the average time needed was longer (9.41 secs), compared to the threshold value of 0.05, which only required 4.96 secs. Therefore, the distinction on the average time was 4.45 secs.

Cancer dataset The second test used the Cancer dataset, which had 569 instances with 31 features. The average value of the highest test accuracy on this dataset was generated by a threshold of 0.05, which was 96.63% with 26 selected features. The CBFS further produced 12 selected features with an average accuracy of 95.79%. Moreover, the proposed threshold had an average accuracy of 94.39%, by using 15 features, the average value is 0.4008 and a standard deviation of 0.2384. However, in terms of the average time required, the proposed threshold was superior to the two methods being compared, which was 0.07 secs.

Furthermore, during the trial of using the transformed Cancer dataset and a threshold of 0.05, the average accuracy value had the highest rate of 96.68%, with 26 selected features. Also, during the trial test for the average time required, the proposed threshold had the same result as the CBFS, which was 0.08 secs with 15 selected features. Therefore, the average accuracy value was only 94.41%, with a standard deviation of 0.2385.

Contraceptive method dataset The third trial was carried out on the Contraceptive Method dataset having 1,473 instances, with 9 features. By using the proposed threshold, a standard deviation value of 0.0324, the average value is 0.0383 was obtained, with four selected features. Furthermore, the average accuracy value generated at this

threshold was higher (51.64%), and outperforming the rate generated by the CBFS and 0.05. However, for the execution of the average time required, the three methods being compared had the same result, which was 0.25 secs.

Also, the trial using the proposed threshold yielded the highest average accuracy value (51.74%), wcompared to the CBFS algorithm with 0.05, through the use of only 4 features and a standard deviation rate of 0.0342. The average time required for execution was faster when using 0.05, compared to the proposed CBFS algorithm and threshold, which only took 0.21 secs.

Dermatology dataset The fourth trial was conducted on the Dermatology dataset, which had 366 instances with 33 features. By using the proposed threshold average accuracy value was higher compared to that of the CBFS method and 0.05. The average value is 0.4205, a standard deviation of 0.2363with 26 selected features. However, for the execution of the average time required, the proposed threshold method only took 0.04 secs. The trials conducted on the Dermatology group were different from those carried out on the other datasets. The Dermatology dataset had a missing value, making it unable to be directly transformed using FFT and IFFT. Therefore, in this test, all missing values were filled with a value of 0 (zero), prior to transformation.

After the transformation, tests were carried out to obtain the average time and accuracy value, as required. Moreover, the CBFS produced the highest average accuracy value (97.70%), compared to the other two methods. Also, in terms of the average time required, the 0.05 and proposed threshold, had the same result of 0.07 secs with a standard deviation of 0.2359.

Divorce dataset The fifth test was carried out on the Divorce dataset, which had 170 instances with 54 features. By using the selected 0.05 and the proposed threshold, the result of the average accuracy value was the same (97.65%), with a standard deviation of 0.1896 and a mean value is 0.6559. Also, both the 0.05 and proposed threshold had selected features of 54 & 52, respectively. However, the average time required was less at the proposed threshold, which was only 0.02 secs.

During the tests on the transformed Divorce dataset, the average accuracy value using the 0.05 threshold, was the highest (97.71%), with the required time being the same as that of the CBFS, which was 0.02 secs. Therefore, as regards the proposed threshold, the average accuracy value had a slight difference with 0.05, which was 97.65% with 53 selected features, and a standard deviation of 0.1920.

Electrical grid dataset The sixth trial was carried out on the Electrical Grid dataset, which had 10,000 instances with 14 features. By using the CBFS, it was discovered that the selected features were 9, with an average accuracy value of 100%. Also, for selection by setting 0.05 and the proposed threshold value, the average accuracy rate was the same (100%), the mean value is 0.1009 with a standard deviation of 0.2546. Moreover, 5 features were selected for the 0.05, while the proposed threshold value had to settle for 1. However, the average time was faster than using the method with the proposed threshold, which was at 0.17 secs.

The tests on the transformed Electrical Grid dataset discovered that, by using the CBFS, the average accuracy value was higher than the other two methods, having 85.64% with 9 selected features. The average time required was less by using a threshold of 0.05, which was 2.57 secs. Therefore, the average time needed for the proposed threshold was slightly longer (3.44 secs), with a standard deviation of 0.0334.

CNAE-9 dataset The seventh test was conducted on the CNAE-9 dataset, which had 1,080 instances with 857 features. The mean value of accuracy of the proposed method is 88.05%, with 65 features selected, a mean value of 0.0121, and a standard deviation of 0.0402. The average accuracy value is higher than the CBFS and a threshold of 0.05. However, the average time needed was less when using the CBFS algorithm (0.27 secs), compared to the proposed threshold. Therefore, there was a difference of 0.31 secs less, compared to the average time required by the proposed threshold.

The tests on the transformed CNAE-9 dataset observed that, by using 0.05, the average accuracy value produced was higher (90.69%), compared to the CBFS algorithm and the proposed threshold, with 57 selected features. The average accuracy value was slightly higher than that of the proposed threshold, with the difference being 0.2% with a standard deviation of 0.0402. Therefore, the average time needed was less by using the CBFS algorithm, which was only 0.27 secs.

Urban land cover dataset The eighth trial was carried out on the Urban Land Cover dataset, having 168 instances and 148 features. By using the CBFS, the average curation value was 87.68%, with the number of features at 148. Also, in terms of the average time needed, this method had a value of 0.06 secs. However, the mean accuracy value of the proposed threshold is 84.76%, with 57 features selected, and the average required time is 0.07 s. The mean value is 0.4883, and the standard deviation is 0.4536.

The tests on the transformed Urban Land Cover dataset observed that, the average accuracy value at the 0.05 and proposed threshold had the same rate (69.73%), with 178 selected features, and a standard deviation of 0.0078. Therefore, as regards the average time required, the CBFS had a lesser value (21.52 secs), compared to the 0.05 and proposed threshold.

Epilepsy dataset The ninth trial was conducted on the Epilepsy dataset, which had 11,500 instances and 179 features. At the 0.05 and the proposed threshold, the average accuracy value had the same result of 69.60% with 178 features selected, the mean value is 0.2939, and a standard deviation of 0.0078. Moreover, the average accuracy value was observed to be higher than the Correlation-Base Feature Selection. However, for the average time required, the CBFS method had a lesser value than the two compared methods, which was 15.72 secs.

The tests on the transformed Epilepsy dataset discovered that, the average accuracy value at the 0.05 and proposed threshold was higher than the CBFS algorithm, at 69.84% with 178 features selected, and a standard deviation value of 0.0078. Therefore, the average time needed was less by using the CBFS algorithm, which was only 18.20 secs.

SCADI dataset The tenth trial was carried out on the SCADI dataset, which had 70 instances and 206 features. In the Correlation-Base Feature Selection, 19 features were

selected with an average accuracy of 84.14%. The average accuracy value is better than that generated by the proposed threshold, namely 83.86% with 64 selected features, an average value of 0.1793, and a standard deviation of 0.2118. However, for the average time needed at the proposed threshold, the result was faster, taking only 0.01 secs.

The tests on the transformed SCADI dataset observed that the average accuracy value when using the CBFS, was higher than that of the other two methods, which was 85.86% with 16 features selected. Therefore, the average time required for Correlation-Base Feature Selection, 0.05 and the proposed threshold, was the same (0.02 secs).

Accuracy and speed on dataset transformation

Tests on the transformed features using the proposed threshold found that the average accuracy value is no better than the original or transformed dataset. It is only in the Electrical Grid dataset that the average accuracy of the features transformed using the proposed threshold is similar to other methods by 100%. Meanwhile, the transformed feature with the proposed threshold of 90% of the trial results yields a less average time.

The features elimination

In this study, the elimination of the feature in each method is quite diverse.

- In the CBFS method, it experienced a reduction between 0.56 and 96.73%,
- 0.56–93.35% for IG with a threshold of 0.05,
- 0.56–92.86% with the proposed threshold.

The feature elimination for the IG method with the proposed threshold on the transformed features experienced a high reduction. The reduction rate was 7.14% to 99.44% for feature transformation using imaginary numbers and 7.14% to 99.44% without imaginary numbers. This means the highest reduction is in the features transformed and uses the IG method with the proposed threshold.

Discussion

Based on the results of the trials conducted using K-Cross Fold Validation, with a value of $K = 10$, the following was obtained.

Average accuracy value

Original dataset

1. The proposed threshold was compared with the Correlation-Base Feature Selection,

   The average trial accuracy using the original dataset showed that 60% of the proposed threshold method produced higher parameters than the CBFS algorithm, with 10% having the same rates.

2. The proposed threshold was compared with the method of 0.05,
   The average trial accuracy using the original dataset showed that 50% of the proposed threshold method produced higher parameters than 0.05, with 30% having the same rates.
3. Comparison of the proposed threshold, 0.05, and Correlation-Base Feature Selection,
   The average accuracy of these 3 methods showed that 40% of the proposed threshold yielded higher parameters than the other two techniques. Also, the 30% average accuracy value of the three techniques all had the same rate. Therefore, the 70% average accuracy value using the proposed threshold method had better rates.
4. Value Standard Deviation
   Seven of the ten datasets have a standard deviation value less than or equal to (<=) the mean value and four of these datasets have an average accuracy value that is better than the Correlation-Base Feature Selection method and the Information Gain method with a threshold of 0.05. Meanwhile, two datasets have the same average accuracy value. Therefore, 85.71% of the dataset has a standard deviation value that is smaller or equal to the mean with a better average accuracy.

Transformation dataset

1. The proposed threshold was compared with the Correlation-Base Feature Selection,

   The trial average accuracy using a transformed dataset showed that 50% of the proposed threshold method produced higher rates than the CBFS algorithm, with 10% possessing similar parameters.
2. The proposed threshold was compared with the method of 0.05,
   The trial average accuracy using the transformed dataset showed that 40% of the proposed threshold method produced higher rates than the 0.05 technique, with 20% having similar parameters.
3. Comparison of the proposed threshold, 0.05, and Correlation-Base Feature Selection,
   The average accuracy of the 3 methods showed that 20% of the proposed threshold yielded higher parameters than the other two techniques, with 10% possessing similar rates. Therefore, 50% average accuracy using the proposed threshold on the transformed data had fewer good parameters than the other methods.

Transformation features

Using the proposed threshold, the feature transformation with FFT yielded a less superior average accuracy value than the original and transformed datasets. In the Electrical Grid dataset, the average accuracy value of the features transformed using the proposed threshold was similar to other methods using the original and transformed data set by 100%.

Average time required

Original dataset

1. The proposed threshold was compared with the Correlation-Base Feature Selection,

   The implementation of the proposed threshold using the original dataset yielded less average execution time than the CBFS algorithm. Furthermore, among the 10 datasets tested, 50% of the proposed threshold required less average execution time than the Correlation-Base Feature Selection, with 10% needing the same period.
2. The proposed threshold was compared with the 0.05 method,
   Among the 10 datasets tested, 70% required less time than the 0.05 method, with 20% requiring the same period.
3. Comparison of the proposed threshold, 0.05, and Correlation-Base Feature Selection,
   Using the proposed threshold, 50% of the tested datasets required a lesser average time than the 0.05 technique and CBFS. Therefore, 10% of these datasets required the same average execution time of 0.25 secs for the Contraceptive Method group.
4. Three datasets have a standard deviation value higher than the Mean (>) value, while two of the three have a faster average time. The other four datasets have standard deviation values that are less than or equal to the mean value, although they have superior accuracy and require a lower meantime.

Transformation dataset

1. The proposed threshold was compared with the Correlation-Base Feature Selection,

   Using the proposed threshold, only 20% of the datasets produced lesser average time, with 30% possessing the same period.
2. The proposed threshold was compared with the 0.05 method,
   At the proposed threshold, 40% of the transformed dataset had a faster average execution time than the 0.05 technique, with 20% needing the same time average.
3. The comparison of the proposed threshold, 0.05, and Correlation-Base Feature Selection,
   Only 10% of the dataset produced a lesser average time than the 0.05 technique and Correlation-Base Feature Selection, with 20% of the groups requiring the same period.

Transformation features

1. The average execution time on the features transformed by FFT, using imaginary values or the proposed threshold, reduced the average time by 70%.
2. Value of Standard Deviation on features transformed with FFT.

Eight out of ten datasets have a standard deviation value higher (>) than the mean value. All datasets need a lower average time than the Correlation-Base Feature Selection method and the Information Gain method with a threshold of 0.05.

The feature elimination

The feature elimination in the proposed threshold method was transformed without imaginary numbers. Therefore, the features reduced from 92.86% to 99.44%, with Fig. 6 used to compare the features eliminated in each method.

The average accuracy and time required for the entire dataset in this experiment are shown in Tables 1 and 2. Similarly, the comparison images are shown in Figs. 2, 3, 4, 5.

Conclusion

From the trials on the original dataset, that the following was concluded.

1. The original dataset's average accuracy value showed that the proposed threshold method produced higher parameters than when using the CBFS and 0.05 techniques. Moreover, 40% of the original dataset result in higher average accuracy values, with 30% having increased rates similar to the 0.05 method. Therefore, 70% of the tested datasets produced higher average values than the other two methods.
2. Using the proposed threshold value, 50% of the datasets used resulted in less average execution time with better mean accuracy than CBFS and 0.05 techniques. Furthermore, 10% of the dataset yielded the same average time with the 0.05 method. There-

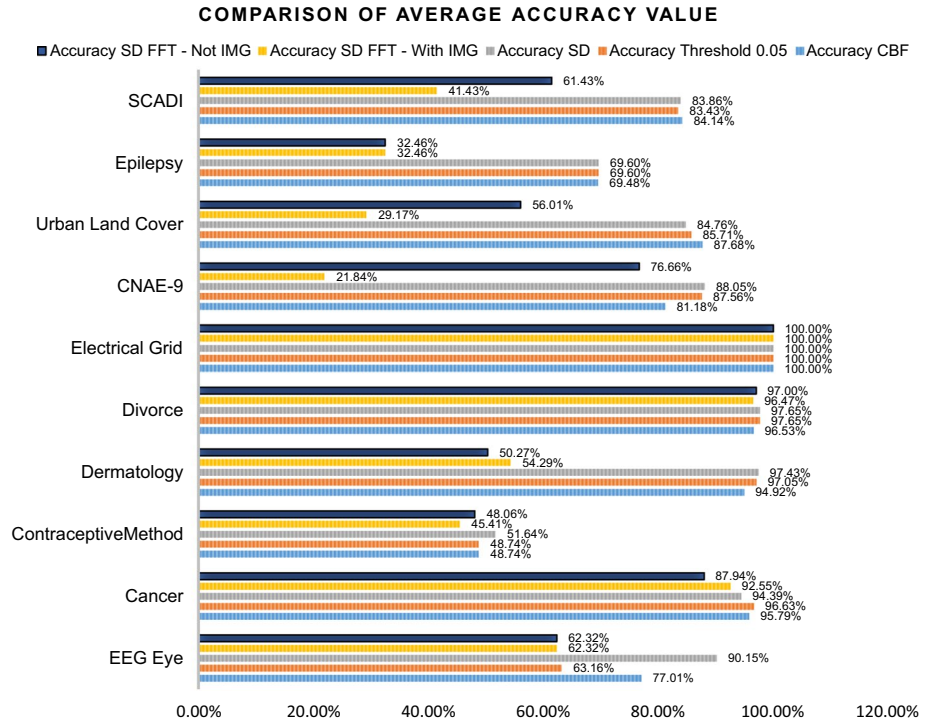

Fig.2 Comparison of average accuracy values

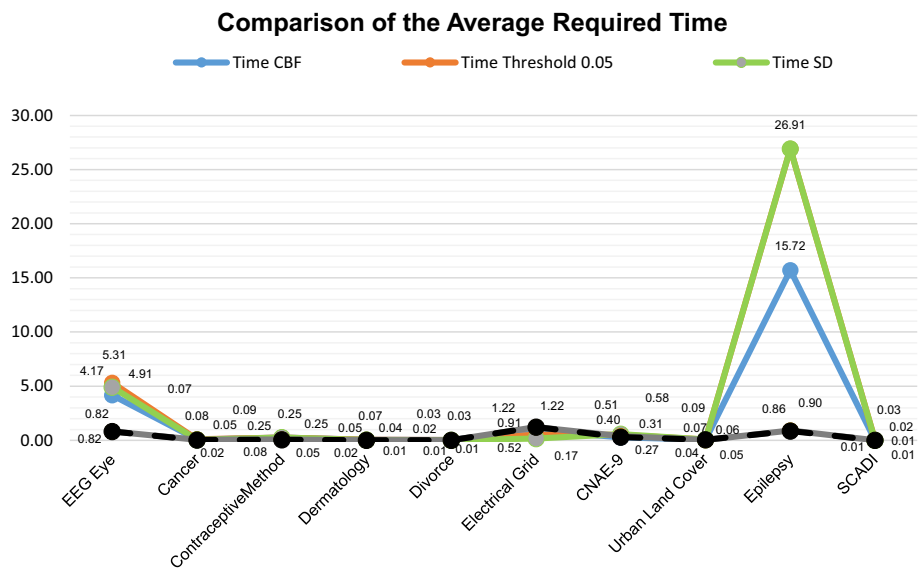

Fig.3 Comparison of the average required time

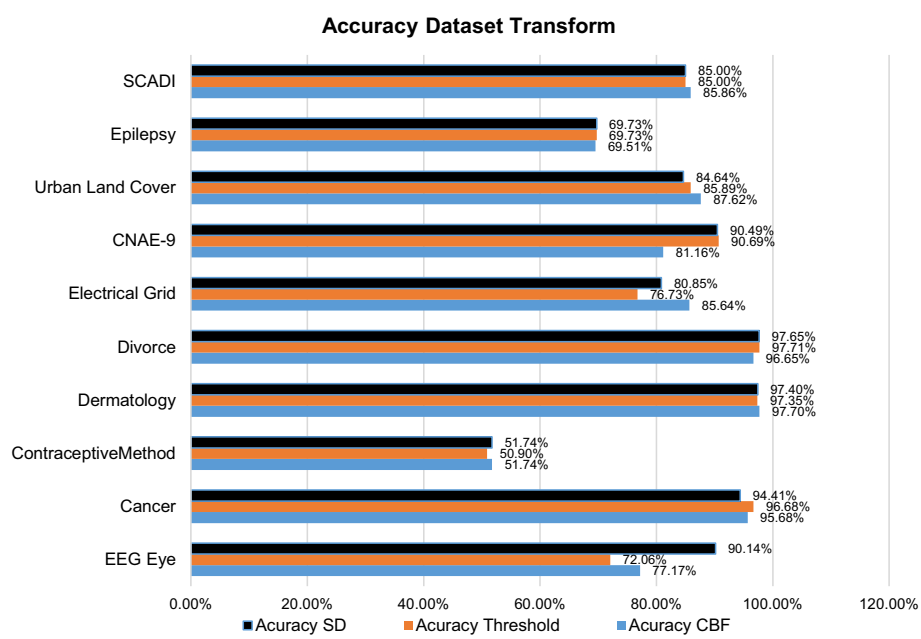

Fig. 4 Comparison for the mean value of transformation dataset accuracy

fore, 60% of the tested dataset led to a shorter average time than the CBFS and 0.05 techniques.

3. Using the transformed feature with the proposed threshold without imaginary numbers, 70% of the trials produced a faster average time, though the average accuracy value obtained is insignificant.

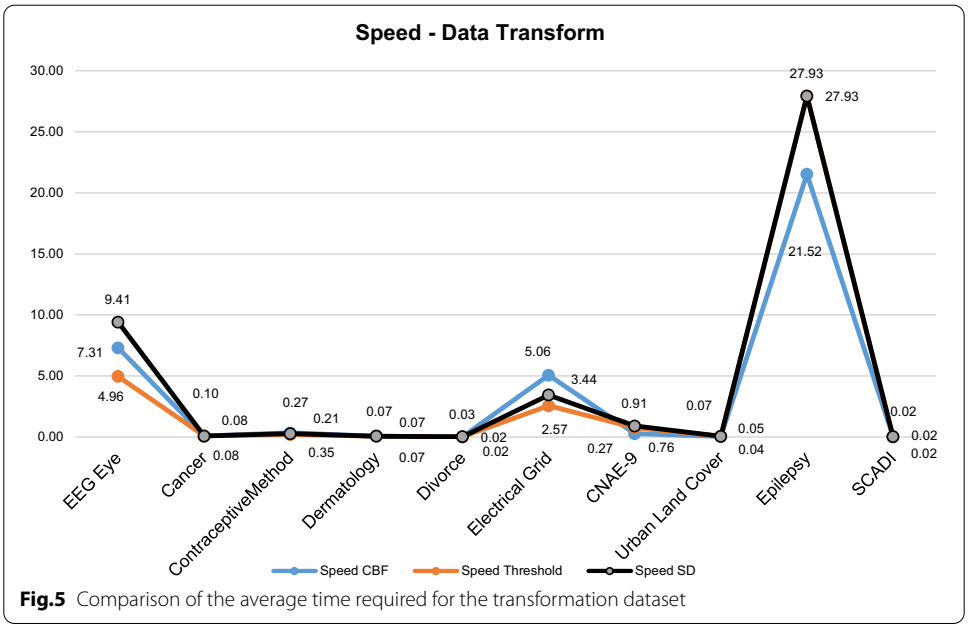

Fig.5 Comparison of the average time required for the transformation dataset

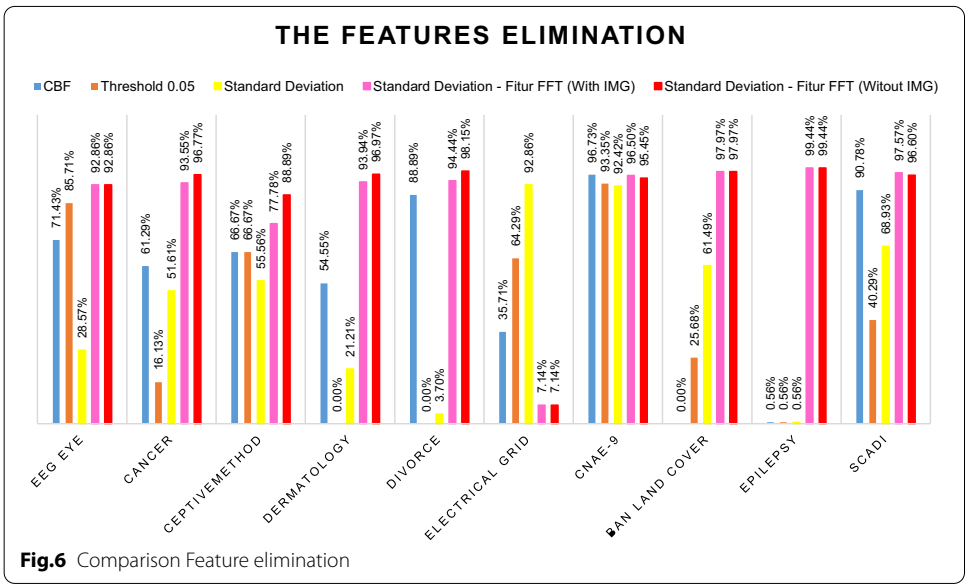

Fig.6 Comparison Feature elimination

4. Based on the experiment, if the standard deviation value is less than or equal to (<=) the mean value, then the accuracy value is superior.
5. The proposed threshold method with transformed features without imaginary numbers significantly reduces features by 92.86% and 99.44%, thereby accelerating the execution process.

The datasets transformed using FFT and IFFT showed the following result;

Table 1 Test results for accuracy and time required

Dataset	CBFS Best First		Threshold 0,05		Treshold based on Standard Deviation		CBFS Best First (FFT)		Threshold 0.05 (FFT)		Treshold based on Standard Deviation (FFT)		Treshold based on Standard Deviation – FFT Fitur with Imaginary		Treshold based on Standard Deviation – FFT Fitur without Imaginary	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
EEG Eye	77.01%	4.17	63.16%	5.31	90.15%	4.91	77.17%	7.31	72.06%	4.96	90.14%	9.41	62.32%	0.82	62.32%	0.82
Cancer	95.79%	0.08	96.63%	0.09	94.39%	0.07	95.68%	0.08	96.68%	0.10	94.41%	0.08	92.55%	0.05	87.94%	0.02
Contraceptive Method	48.74%	0.25	48.74%	0.25	51.64%	0.25	51.74%	0.35	50.90%	0.21	51.74%	0.27	45.41%	0.08	48.06%	0.05
Dermatology	94.92%	0.07	97.05%	0.05	97.43%	0.04	97.70%	0.07	97.35%	0.07	97.40%	0.07	54.29%	0.02	50.27%	0.01
Divorce	96.53%	0.03	97.65%	0.03	97.65%	0.02	96.65%	0.02	97.71%	0.02	97.65%	0.03	96.47%	0.01	97.00%	0.01
Electrical Grid	100.00%	0.91	100.00%	0.52	100.00%	0.17	85.64%	5.06	76.73%	2.57	80.85%	3.44	100%	1.22	100%	1.22
CNAE-9	81.18%	0.27	87.56%	0.51	88.05%	0.58	81.16%	0.27	90.69%	0.76	90.49%	0.91	21.84%	0.40	76.66%	0.31
Urban Land Cover	87.68%	0.06	85.71%	0.09	84.76%	0.07	87.62%	0.04	85.89%	0.07	84.64%	0.05	29.17%	0.04	56.01%	0.05
Epilepsy	69.48%	15.72	69.60%	26.91	69.60%	26.91	69.51%	21.52	69.73%	27.93	69.73%	27.93	32.46%	0.90	32.46%	0.86
SCADI	84.14%	0.02	83.43%	0.03	83.86%	0.01	85.86%	0.02	85.00%	0.02	85.00%	0.02	41.43%	0.01	61.43%	0.01

Table 2 Test results for accuracy and time of the entire transformation dataset

Dataset	Number of Instance	Number of Feature	CBFS Best First			Threshold 0,05			Treshold based on Standard Deviation		
			Number of Feature	Accuracy	Time	Number of Feature	Accuracy	Time	Number of Feature	Accuracy	Time
EEG Eye	14.980	14	4	77.17%	7.31	3	72.06%	4.96	10	90.14%	9.41
Cancer	569	31	12	95.68%	0.08	26	96.68%	0.10	15	94.41%	0.08
ContraceptiveMethod	1.473	9	4	51.74%	0.35	3	50.90%	0.21	4	51.74%	0.27
Dermatology	366	33	15	97.70%	0.07	32	97.35%	0.07	26	97.40%	0.07
Divorce	170	54	6	96.65%	0.02	54	97.71%	0.02	53	97.65%	0.03
Electrical Grid	10.000	14	9	85.64%	5.06	5	76.73%	2.57	7	80.85%	3.44
CNAE-9	1.080	857	28	81.16%	0.27	57	90.69%	0.76	65	90.49%	0.91
Urban Land Cover	168	148	28	87.62%	0.04	110	85.89%	0.07	65	84.64%	0.05
Epilepsy	11.500	179	119	69.51%	21.52	178	69.73%	27.93	178	69.73%	27.93
SCADI	70	206	16	85.86%	0.02	58	85.00%	0.02	58	85.00%	0.02

1. The average accuracy value from the proposed threshold was insignificant compared to the CBFS and 0.05 techniques. The proposed threshold only produced an average accuracy value of 30% higher than the dataset.
2. The proposed threshold yielded an insignificant average time during transformation, with 70% of the transformed dataset taking longer. Therefore, only 30% of the dataset required less time than the 0.05 method and CBFS.
3. Based on the experiment, if the standard deviation value is higher (>) than the mean value, then the time needed is less (faster).
4. The Random Forest execution feature selection on 70% of the transformed dataset increased the average accuracy value between 0.01 and 2.61%. Furthermore, 60% took less time than the original groups, with 10% requiring the same period. Therefore, the difference in the time needed was between 0.01 and 3.05 s.

The trial results showed that several things need to be considered, as follows

Transformations do not need to be used on datasets with incomplete data. Furthermore, pre-processing is required for the data set to be complete. Transformation features can be proposed for further research by combining feature selection and extraction methods such as Principal Component Analysis (PCA), Neural Network, or Singular Value Decomposition (SVD).

The implementation of FFT and IFFT in the dataset needs to be considered, especially in the IG method with the proposed threshold.

Execution time (speed) and accuracy value are inversely proportional to variables, which means that preference is required. If you need to find a superior average accuracy value, you can use IG with a threshold standard deviation using the original data set. Meanwhile, to get is increased average speed using transformation features.

Abbreviations
CBFS: Correlation-Base Feature Selection; FFT: Fast Fourier Transform; IFFT: Inverse Fast Fourier Transform; IG: Information Gain.

Acknowledgements
We would like to thank Institut Teknologi Bandung and Universitas Multimedia Nusantara (UMN) for supporting this research.

Authors' contributions
The author confirms the sole responsibility for this manuscript fully as a sole author for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. The author read and approved the final manuscript.

Funding
Not applicable. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The original dataset used for this study is available in: UCI Machine Learning Repository (www.arsip.lcs.uci.edu/ml). Kaggle Datasets (www.kaggle.com/datasets)

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The author reports no potential conflict of interest.

Received: 14 December 2020 Accepted: 22 May 2021
Published online: 05 June 2021

References

1. Khalid S, Khalil T, Nasreen S A survey of feature selection and feature extraction techniques in machine learning. In: 2014 Science and information conference, London, UK; 2014. p. 372–378. Doi: https://doi.org/10.1109/SAI.2014.6918213.
2. Hira ZM, Gillies DF. A Review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform. 2015;2015:1–13. https://doi.org/10.1155/2015/198363.
3. Corizzo R, Ceci M, Japkowicz N. Anomaly detection and repair for accurate predictions in geo-distributed big data. Big Data Res. 2019;16:18–35. https://doi.org/10.1016/j.bdr.2019.04.001.
4. Corizzo R, Ceci M, Zdravevski E, Japkowicz N. Scalable auto-encoders for gravitational waves detection from time series data. Expert Syst Appl. 2020;151:113378. https://doi.org/10.1016/j.eswa.2020.113378.
5. Zheng K, Li T, Zhang B, Zhang Y, Luo J, Zhou X. Incipient fault feature extraction of rolling bearings using auto-correlation function impulse harmonic to noise ratio index based SVD and teager energy operator. Appl Sci. 2017;7(11):1117. https://doi.org/10.3390/app7111117.
6. Gu Y, Yang X, Peng M, Lin G. Robust weighted SVD-type latent factor models for rating prediction. Expert Syst Appl. 2020;141:112885. https://doi.org/10.1016/j.eswa.2019.112885.
7. Wei G, Zhao J, Feng Y, He A, Yu J. A novel hybrid feature selection method based on dynamic feature importance. Appl Soft Comput. 2020;93:106337. https://doi.org/10.1016/j.asoc.2020.106337.
8. Prasetiyowati MI, Maulidevi NU, Surendro K. The speed and accuracy evaluation of random forest performance by selecting features in the transformation data. In: IEEA 2020: proceedings of the 2020 the 9th international conference on informatics, environment, energy and applications. 2020. p. 125–130. Doi: https://doi.org/10.1145/3386762.3386768.
9. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
10. Ma J, Gao X. A filter-based feature construction and feature selection approach for classification using Genetic Programming. Knowl-Based Syst. 2020;196:105806. https://doi.org/10.1016/j.knosys.2020.105806.
11. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal. 2020;143:106839. https://doi.org/10.1016/j.csda.2019.106839.
12. Thabtah F, Kamalov F, Hammoud S, Shahamiri SR. Least Loss: A simplified filter method for feature selection. Inf Sci. 2020;534:1–15. https://doi.org/10.1016/j.ins.2020.05.017.
13. Samami M, et al. A mixed solution-based high agreement filtering method for class noise detection in binary classification. Phys A. 2020;553:124219. https://doi.org/10.1016/j.physa.2020.124219.
14. Das H, Naik B, Behera HS. A Jaya algorithm based wrapper method for optimal feature selection in supervised classification. J King Saud Univ Comput Inf Sci. 2020. https://doi.org/10.1016/j.jksuci.2020.05.002.
15. González J, Ortega J, Damas M, Martín-Smith P, Gan JQ. A new multi-objective wrapper method for feature selection—accuracy and stability analysis for BCI. Neurocomputing. 2019;333:407–18. https://doi.org/10.1016/j.neucom.2019.01.017.
16. Lu M. Embedded feature selection accounting for unknown data heterogeneity. Expert Syst Appl. 2019;119:350–61. https://doi.org/10.1016/j.eswa.2018.11.006.
17. Zhang P, Gao W. Feature selection considering Uncertainty Change Ratio of the class label. Appl Soft Comput. 2020;95:106537. https://doi.org/10.1016/j.asoc.2020.106537.
18. Elmaizi A, Nhaila H, Sarhrouni E, Hammouch A, Nacir C. A novel information gain based approach for classification and dimensionality reduction of hyperspectral images. Proc Comput Sci. 2019;148:126–34. https://doi.org/10.1016/j.procs.2019.01.016.
19. Jadhav S, He H, Jenkins K. Information gain directed genetic algorithm wrapper feature selection for credit rating. Appl Soft Comput. 2018;69:541–53. https://doi.org/10.1016/j.asoc.2018.04.033.
20. Singer G, Anuar R, Ben-Gal I. A weighted information-gain measure for ordinal classification trees. Expert Syst Appl. 2020;152:113375. https://doi.org/10.1016/j.eswa.2020.113375.
21. Dems˘ar J, Demsar J. Statistical comparisons of classifi ers over multiple data sets. J Mach Learn Res. 2006;7:1–30.
22. Yang Z, et al. Robust discriminant feature selection via joint L 2, 1 -norm distance minimization and maximization. Knowl Based Syst. 2020. https://doi.org/10.1016/j.knosys.2020.106090.
23. Tsai C-F, Sung Y-T. Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. Knowl-Based Syst. 2020;203:106097. https://doi.org/10.1016/j.knosys.2020.106097.
24. Leo B. Bagging predictors. Mach Learn. 1996;24(2):123–40.
25. Herff C, Krusienski DJ. Extracting features from time series. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of clinical data science. Cham: Springer International Publishing; 2019. p. 85–100.
26. Li M, Chen W. FFT-based deep feature learning method for EEG classification. Biomed Signal Process Control. 2021;66:102492. https://doi.org/10.1016/j.bspc.2021.102492.
27. Seco GBS, Gerhardt GJL, Biazotti AA, Molan AL, Schönwald SV, Rybarczyk-Filho JL. EEG alpha rhythm detection on a portable device. Biomed Signal Process Control. 2019;52:97–102. https://doi.org/10.1016/j.bspc.2019.03.014
28. Ansari MF, Edla DR, Dodia S, Kuppili V. Brain-computer interface for wheelchair control operations: an approach based on fast fourier transform and on-line sequential extreme learning machine. Clin Epidemiol Global Health. 2019;7(3):274–8. https://doi.org/10.1016/j.cegh.2018.10.007.
29. Hosseini S, Roshani GH, Setayeshi S. Precise gamma based two-phase flow meter using frequency feature extraction and only one detector. Flow Meas Instrum. 2020;72:101693. https://doi.org/10.1016/j.flowmeasinst.2020.101693.

30. Gowid S, Dixon R, Ghani S. A novel robust automated FFT-based segmentation and features selection algorithm for acoustic emission condition based monitoring systems. Appl Acoust. 2015;88:66–74. https://doi.org/10.1016/j.apacoust.2014.08.007.
31. Prasetiyowati MI, Maulidevi NU, Surendro K. Feature selection to increase the random forest method performance on high dimensional data. Int J Adv Intell Inf. 2020;6(3):10.
32. Lei S. A feature selection method based on information gain and genetic algorithm. In: 2012 international conference on computer science and electronics engineering, Hangzhou, Zhejiang, China; 2012. p. 355–358. Doi: https://doi.org/10.1109/ICCSEE.2012.97.
33. Genuer R, Poggi J, Tuleau-malot C, Villa-vialaneix N. Random Forests for Big Data. Big Data Res. 2017;1:1–19. https://doi.org/10.1016/j.bdr.2017.07.003.
34. Breiman LEO. Random forests. Netherlands: Kluwer Academic Publishers; 2001.
35. Ye Y, Wu Q, Zhexue Huang J, Ng MK, Li X. Stratified sampling for feature subspace selection in random forests for high dimensional data. Pattern Recogn. 2013;46(3):769–87. https://doi.org/10.1016/j.patcog.2012.09.005.
36. Chen M-Y, Chen B-T. Online fuzzy time series analysis based on entropy discretization and a Fast Fourier Transform. Appl Soft Comput. 2014;14:156–66. https://doi.org/10.1016/j.asoc.2013.07.024.
37. Ghaderi H, Kabiri P. Fourier transform and correlation-based feature selection for fault detection of automobile engines. In: The 16th CSI international symposium on artificial intelligence and signal processing (AISP 2012), Shiraz, Fars, Iran; 2012. p. 514–519. Doi: https://doi.org/10.1109/AISP.2012.6313801.
38. Sim J, Lee JS, Kwon O. Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. Math Prob Eng. 2015;2015:1-14. https://doi.org/10.1155/2015/538613
39. Ichikawa M, Hosono A, Tamai Y, Watanabe M, Shibata K, Tsujimura S, Oka K, Fujita H, Okamoto N, Kamiya M, Kondo F, Wakabayashi R, Noguchi T, Isomura T, Imaeda N, Goto C, Yamada T, Suzuki S. Handling missing data in an FFQ: multiple imputation and nutrient intake estimates. Public Health Nutr. 2019;22(8):1351–1360. https://doi.org/10.1017/S1368980019000168
40. Hening D, Koonce DA. Missing data imputation method comparison in ohio university student retention database, p. 10.
41. Dua D, Graff C. UCI machine learning repository. University of California, School of Information and Computer Science. [Online]. http://archive.ics.uci.edu/ml.
42. "Breast Cancer Wisconsin (Diagnostic) Data Set Predict whether the cancer is benign or malignant." [Online]. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.
43. Yöntem MK, Ilhan T. Divorce prediction using correlation based feature selection and artificial neural networks. Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi. 2019.
44. Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. Phys Rev E. 2001;64(6):061907. https://doi.org/10.1103/PhysRevE.64.061907.
45. Zarchi MS, Fatemi Bushehri SMM, Dehghanizadeh M. SCADI: A standard dataset for self-care problems classification of children with physical and motor disability. Int J Med Inf. 2018;114:81–7. https://doi.org/10.1016/j.ijmedinf.2018.03.003.
46. Fatemi Bushehri SMM, Zarchi MS. An expert model for self-care problems classification using probabilistic neural network and feature selection approach. Appl Soft Comput. 2019;82:105545. https://doi.org/10.1016/j.asoc.2019.105545.
47. Johnson B, Xie Z. Classifying a high resolution image of an urban area using super-object information. ISPRS J Photogramm Remote Sens. 2013;83:40–9. https://doi.org/10.1016/j.isprsjprs.2013.05.008.
48. Johnson B. High-resolution urban land-cover classification using a competitive multi-scale object-based approach. Remote Sens Lett. 2013;4(2):131–40. https://doi.org/10.1080/2150704X.2012.705440.

Publisher's Note