Determining Threshold Value on Information Gain Feature Selection to Increase Speed and Prediction Accuracy of Random Forest

Maria Irmina Prasetiyowati (maria@umn.ac.id)
Institut Teknologi Bandung https://orcid.org/0000-0003-1815-7068
Nur Ulfa Maulidevi
Institut Teknologi Bandung
Kridanto Surendro
Institut Teknologi Bandung

Research

Keywords: Threshold, standard deviation, accuracy, time, Random Forest

Posted Date: December 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-132775/v1

Determining Threshold Value on Information Gain Feature Selection to Increase Speed and Prediction Accuracy of Random Forest

Maria Irmina Prasetiyowati[1]*, Nur Ulfa Maulidevi[2], Kridanto Surendro[3]

School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia

**Corresponding author: maria@umn.ac.id*

Abstract

Feature selection is a preprocessing technique aims to remove the unnecessary features and speed up the algorithm's work process. One of the feature selection techniques is by calculating the information gain value of each feature in a dataset. From the information gain value obtained, then the determined threshold value will be used to make feature selection. Generally, the threshold value is used freely, or using a value of 0.05. This study proposed the determination of the threshold value using the standard deviation of the information gain value generated by each feature in the dataset. The determination of this threshold value was tested on ten original datasets and datasets that had been transformed by FFT and IFFT, then classified using Random Forest. The results of the average value of accuracy and the average time required from the Random Forest classification using the proposed threshold value are better compared to the results of feature selection with a threshold value of 0.05 and the Correlation-Base Feature Selection algorithm. Likewise, the result of the average accuracy value of the proposed threshold using a transformed dataset in terms are better than the threshold value of 0.05 and the Correlation-Base Feature Selection algorithm. However, the calculation results for the average time required are higher (slower).

Keywords: Threshold, standard deviation, accuracy, time, Random Forest

1. Introduction

Feature selection is a technique that is often used in pre-processing that will affect the performance of a model. The purpose of feature selection is to remove the excessive and unimportant features [1]. Apart from eliminating unnecessary features, feature selection also aims to speed up the algorithm's work process [2]. There are several

ways to perform feature selection [3], which are broadly distinguished into three main types, those are Filter [4–7], Wrapper [8], [9] and Embedded [10]. Filter method is a method that evaluates a subset of features with predefined evaluation criteria and does not depend on any grouping [11]. One of the well-known types of filter feature selection is Information Gain (IG). IG is a technique that performs feature weighting scoring using the maximum entropy value. Although IG is a basic technique in feature selection, which is still open big opportunities for further research and development in this field. The following three studies showed that IG was still used for feature selection. Elmaizi, in his paper entitled a novel IG based approach for classification and dimensionality reduction of hyperspectral images proposed a new approach based on IG for hyperspectral image classification and dimensional reduction [12]. This is in contrast to the research conducted by Jadhav et al, who proposed feature selection by ranking features based on IG. The algorithm he used is called the Gain Directed Feature Selection algorithm (IGDFS). IGDS uses three algorithms, namely Support Vector Machine, KNN and Naïve Bayes [13]. Meanwhile, Singer in his paper entitled a weighted information-gain measure for ordinal classification trees proposed a model that defines proportionally weighted entropy, which is called Weighted Information-Gain (WIGR) [14].

At IG all the features in the dataset will be calculated before some its features will be selected later on. The selected feature is defined by a value limit called a threshold or (arbitrary cutoff). The threshold value is set freely, as needed. In general, the threshold value used is 0.05 [15],[16]. Whereas, Tsai and Sung in their research method in a paper entitled ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches, used the calculation of the frequency of each feature and used the average to get the threshold value as a subset of final features [17]. Based on these studies, the researcher proposed the determination of the taken threshold value based on the standard deviation of the IG value. The weighting result for each feature was calculated and the threshold value determination by using a standard deviation.

Fast Fourier Transform (FFT) is an algorithm that can be applied to increase execution speed. This method divides the original vector into two parts, calculates the FFT of each part, and then combines them. This is done recursively. Hamid used the FFT algorithm to improve the classification results using feature extraction and signal processing in his research [18]. In addition, Herff also wrote in his book, entitled Extracting Features from Time Series, explaining that one of the time series data that is often processed and collected in sequential time

series is the clinical data. This data can be presented in a continuous wave form, so that it can be presented briefly by using the FFT algorithm [19]. Other researchers also used the FFT algorithm to perform feature extraction [20], [21]. Pratama, in his research stated that real and imaginary numbers resulted from the FFT calculation can be used for the feature selection algorithm [22]. However, the relationship between these two values can be lost, and does not maintain dependence between the two attributes, especially in feature selection algorithms [23]. FFT and IFFT can also be applied to a dataset. One example is the research conducted by Prasetiyowati. Her research, entitled the Speed and Accuracy Evaluation of Random Forest Performance by Selecting Features in the Transformation Data, found that the use of a transformed dataset generate the better average value of accuracy and time required than the original dataset [2].

2. Data and Workflow

2.1 Data Collection

In the early stages, the researcher selected several datasets available in the UCI Machine Learning Repository [24], those are the EEG Eye dataset, Cancer [25], Contraceptive Method, Dermatology, Divorce [26], Electrical Grid, CNAE-9, Urban Lan Cover[27], [28], Epilepsy [29] and SCADI [30],[31].

2.2 Workflow

The researcher divided this research into two stages. The first stage was conducting experiments using the original data in the dataset, while the second stage was using a dataset that had been transformed using FFT and returned with IFFT. In general, the stages of the research can be seen in Figure 1 and Figure 2.

The research started from collecting ten existing datasets, which were obtained from UCI and Kaggle. The next step was to check whether the dataset needs to be transformed or not. If transformation is required, the existing dataset is transformed using the FFT algorithm and then returned again using IFFT. After that, the calculation of the accuracy value and time required for execution of each dataset was carried out. However, if the dataset does not need to be transformed, then immediately calculate the accuracy value and time required for Random Forest prediction.

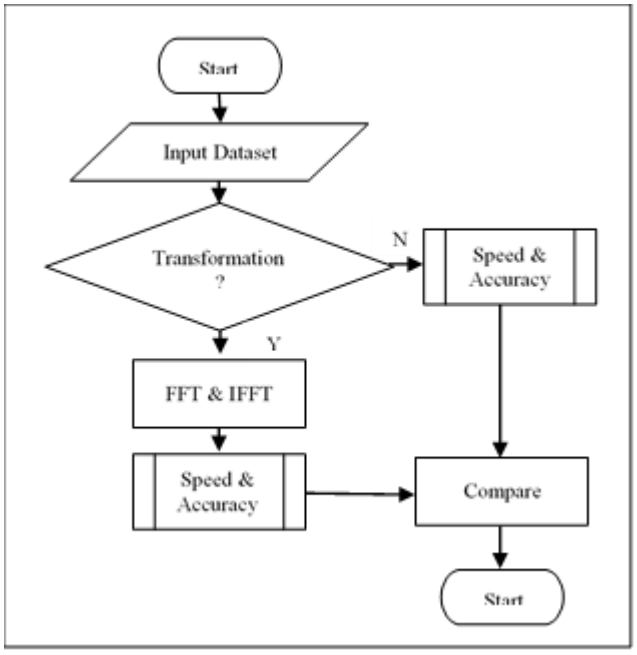

Fig.1 Research Stages

The calculation of the value of accuracy and time required requires the following steps:

1. Collecting the dataset.
2. Selecting the IG feature through the Ranker method, using the Weka machine learning tools version 3.9.2.

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

$$\text{gain}\,(y, A) = entropy(y) \sum_{Cenilai(A)} \frac{Yc}{y} entropy(yc)$$

3. Calculating the standard deviation of the IG for each feature.

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

   After the standard deviation was found, this value was used as the threshold for the dataset.

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

The random forest prediction process was carried out using 10 randomly entered seeds. The selected seed values were 1, 33, 57, 70, 153, 251, 300, 457, 505, and 700. Then, these steps were carried out for each dataset.

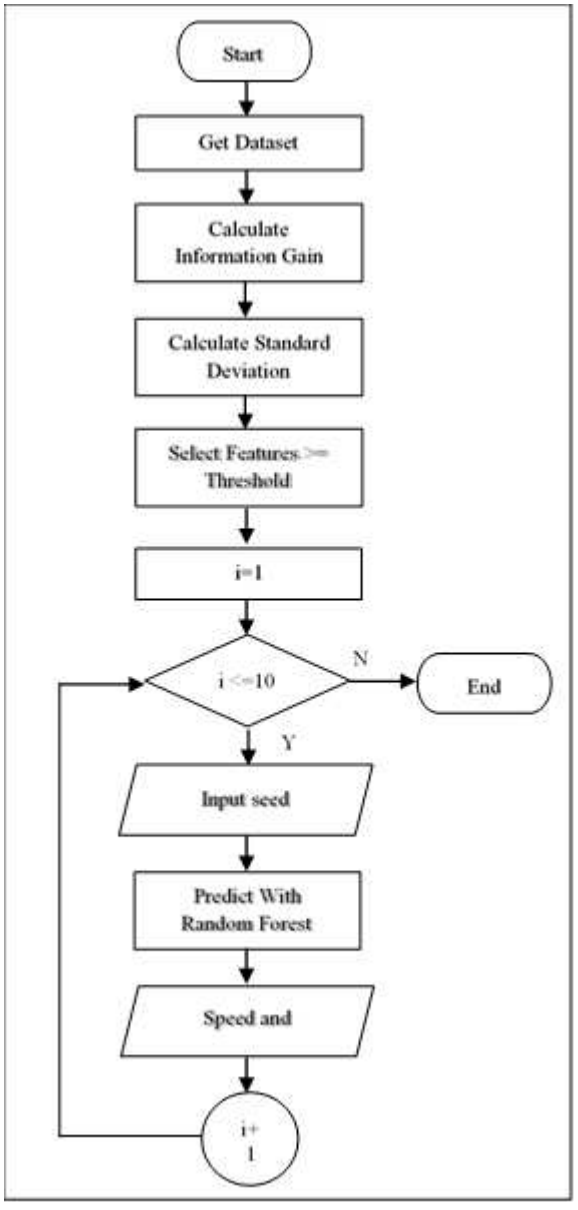

Fig.2 Fowchart Speed and Accuracy

After making predictions using the Random Forest algorithm, the test results with the proposed threshold were compared with the prediction results using the Correlation-Base Feature Selection (CBFS) and the threshold of 0.05. The comparisons is done by using both the original dataset and the transformed dataset.

3. Methods

3.1. Information Gain

Information Gain is an entropy-based feature selection method [32], by calculating from the output data or dependent y features grouped by feature A. This is denoted as gain (y, A). Information Gain (y, A) is denoted as

$$gain(y, A) = entropy(y) \sum_{C \in value(A)} \frac{Yc}{y} entropy(yc)$$

The value (A) is all possible values of attribute A, and Yc is the subset of y where A has the value c. The rule of equation (1) is the total entropy of y, and the next rule is the entropy after segregating the data based on feature A.

3.2 Standard Deviation

One way to determine the diversity of a data group is by reducing the value of the data by the mean of the data group, then add the results. This method is known as standard deviation, which is a description of how much the difference between the measured data or the distribution of a group of data against the average value. The data group referred to in this study is the IG value for each feature in a dataset. Equation (2) can be applied to features that have calculated the IG value in a dataset.

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

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

3.3 Fast Fourier Transform And Inverse Fast Fourier Transform

An effective way to convert time domain signals to frequency domination is by using Fast Fourier Transform (FFT) [33]. The converted data can be returned to the original domain using the Inverse Fast Fourier Transform (IFFT) algorithm. The test of the transformed data using FFT and IFFT is applied to high-dimensional and regular datasets (which have less than 100 features).

The equation for FFT can be seen in Equation 3 and Equation 4 for IFFT.

$$X[k] = \sum_{n=0}^{N-1} X[n] W_N^{kn}, \quad k=0,1,\ldots N-1 \qquad (3)$$

$$X[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] W_N^{-kn}, \quad n=0,1,\ldots N-1 \qquad (4)$$

4 Result

This research use the Random Forest method with K-Cross Fold Validation, with a value of K = 10. Each dataset use ten randomly generated seed values. The ten seed values entered were 1, 33, 57, 70, 153, 251, 300, 457, 505, and 700. Tests were carried out using the original dataset and the FFT - IFFT dataset transformed. The feature selection technique used was the IG. The IG value of each feature obtained was used to calculate the standard deviation value as a threshold determination. The result of feature selection was applied to the dataset used. Then after that the dataset was analyzed using Random Forest. The average accuracy and time required using the proposed threshold were compared with the average accuracy and time values generated by the Correlation-based Feature Selection method and the threshold with a value of 0.05, using either the original dataset or the transformed dataset.

4.1 Accuracy and Speed on EEG Dataset

The first test was conducted on the EEG Eye dataset which has 14,980 instances with 14 features. Through the IG using the proposed threshold, the standard deviation value was 0.0171, so that 10 features that have an IG value above 0.0171 were selected. The resulting average accuracy value was higher, which is 90.15% outperforming the average value generated by the CBFS and threshold 0.05. However, the average execution time was faster using the Correlation-Base Feature Selection. The distinction of the average time is 0.74 seconds.

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

4.2 Accuracy and Speed on 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 is 96.63%, with 26 selected

features. The CBFS produces 12 selected features with an average accuracy of 95.79%. Meanwhile, the proposed threshold had an average accuracy of 94.39% using 15 features 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 is 0.07 seconds.

Meanwhile, the trial using the transformed cancer dataset and using a threshold of 0.05, the average accuracy value had the highest value of 96.68%, with 26 selected features. Whereas for the average time required, the proposed threshold had the same average time as the Correlation-Base Feature Selection, which is 0.08 seconds with 15 selected features. Meanwhile, the average accuracy value was only 94.41% with a standard deviation of 0.2385.

4.3 Accuracy and Speed on Contraceptive Method Dataset

The third trial was carried out on the Contraceptive Method dataset which had 1,473 instances with 9 features. By using the proposed threshold, a standard deviation value of 0.0324 was obtained, so that there were 4 features selected that had an IG value above 0.0324. The average accuracy value generated at this threshold was higher, which is 51.64% outperforming the average accuracy value generated by the CBFS and threshold of 0.05. Meanwhile, for the average time required for the execution, the three methods being compared had the same average time, which is 0.25 seconds.

Meanwhile, the trial using the proposed threshold yields the highest average accuracy value compared to the CBFS algorithm and a threshold of 0.05, which is 51.74%, using only 4 features and a standard deviation value of 0.0342. Meanwhile, the average time required for execution was faster using a threshold of 0.05, compared to the proposed CBFS algorithm and threshold, which only took 0.21 seconds.

4.4 Accuracy and Speed on Dermatology Dataset.

The fourth trial was conducted on the Dermatology dataset which had 366 instances with 33 features. By using the proposed threshold, the average accuracy value was higher than the CBFS method and a threshold of 0.05, which is 97.43% with a standard deviation of 0.2363, and 26 selected features. Likewise with the average time required, the proposed threshold method only took 0.04 seconds.

Meanwhile, the trials conducted on the Dermatology dataset were different from trials with other datasets. The Dermatology dataset had a missing value so that it could not be directly transformed using FFT and IFFT. In this test, all missing values were filled with a value of 0 (zero), prior to transformation.

After the data was transformed, then tests were carried out to obtain the average accuracy value and the average time required. CBFS produced the highest average accuracy value compared to the other two methods, which is 97.70%. In terms of the average time required, the threshold of 0.05 and the proposed threshold had the same average time of 0.07 seconds with a standard deviation of 0.2359.

4.5 Accuracy and Speed on Divorce Dataset.

The fifth test was conducted on the Divorce dataset which had 170 instances with 54 features. By using the feature selection that sets a threshold of 0.05 and the proposed threshold value, the average accuracy value had the same, which is 97.65% with 54 features selected for the 0.05 threshold, and 52 features for the proposed threshold value with a standard deviation of 0.1896. Meanwhile, the average time required was less at the proposed threshold, which is only 0.02 seconds.

Meanwhile, the tests on the transformed Divorce dataset, the average accuracy value using the 0.05 threshold had the highest value, which is 97.71% and the average time required for the execution was the same as the Correlation-Base Feature Selection, which is 0.02 seconds. As for the proposed threshold, the average accuracy value had a slight difference with the threshold of 0.05, which is 97.65% with 53 selected features and a standard deviation of 0.1920.

4.6 Accuracy and Speed on 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 Correlation-Base Feature Selection, among the 14 features in the dataset, it was found that the selected features were 9 with an average accuracy value of 100%. Likewise for feature selection by setting a threshold of 0.05 and the proposed threshold value, the average accuracy value was the same, which is 100% with 5 features selected for the 0.05 threshold and 1 feature for the proposed threshold value with a standard deviation of 0.2546. However, the average time was faster than using the method with the proposed threshold, which is 0.17 seconds.

Tests on the transformed Electrical Grid dataset found that by using the Correlation-Base Feature Selection, the average accuracy value was higher than the two methods, which was 85.64%, using 9 selected features. Meanwhile, the average execution time required was less by using a threshold of 0.05, namely 2.57 seconds. Meanwhile, the average time needed for the proposed threshold was slightly longer, which is 3.44 seconds with a standard deviation of 0.0334.

4.7 Accuracy and Speed on CNAE-9 Dataset.

The seventh test was carried out on the CNAE-9 dataset which had 1,080 instances with 857 features. By using the proposed threshold, the average accuracy value was higher than the CBFS and a threshold of 0.05, which is 88.05% with 65 selected features, and a standard deviation of 0.0402. Meanwhile, the average time needed was less using the CBFS algorithm compared to the proposed threshold, which is 0.27 seconds. There was a difference of 0.31 seconds less than the average time required by the proposed threshold.

Tests on the transformed CNAE-9 dataset found that by using a threshold of 0.05, the average accuracy value produced was higher than the CBFS algorithm and the proposed threshold, which is 90.69% with 57 selected features. The average accuracy value was slightly higher than the average value of the proposed threshold accuracy. The difference was only 0.2% with a standard deviation of 0.0402. Meanwhile, the average time needed was less using the CBFS algorithm, which is only 0.27 seconds.

4.8 Accuracy and Speed on Urban Land Cover Dataset.

The eighth trial was carried out on the Urban Land Cover dataset which had 168 instances and 148 features. By using the Correlation-Base Feature Selection, the average curation value was 87.68%, with the number of features 148. Likewise with the average time needed. This method has an average time less, which is 0.06 seconds. While the average proposed threshold accuracy value is 84.76%, with 57 selected features and a standard deviation of 0.4536, with an average time required of 0.07.

Tests on the transformed Epilepsy dataset found that the average accuracy value at the 0.05 threshold and the proposed threshold had the same value, namely 69.73% with 178 selected features and a standard deviation of 0.0078. As for the average time required, the CBFS required less time than the 0.05 threshold and the proposed threshold, namely 21.52 seconds

4.9 Accuracy and Speed on Epilepsy Dataset.

The ninth trial was carried out on the Epilepsy dataset which had 11,500 instances and 179 features. At the 0.05 threshold and the proposed threshold, the average accuracy value had the same value of 69.60% with 178 features, with a standard deviation of 0.0078. The average accuracy value was higher than the Correlation-Base Feature Selection. However, for the average time required, the CBFS method had an average time less than the two compared methods, which is 15.72 seconds.

Tests on the transformed Epilepsy dataset found that the average accuracy value at the 0.05 threshold and the proposed threshold was higher than the CBFS algorithm, which is 69.84% using 178 features with a standard deviation value of 0.0078. Meanwhile, the average time needed was less using the CBFS algorithm, which is only 18.20 seconds.

3.2.9 Accuracy and Speed on 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 was better than the average accuracy value generated by the proposed threshold, which is 83.86% using 64 features and a standard deviation of 0.2118. Meanwhile, for the average time needed, at the proposed threshold the average time needed was faster, only 0.01 seconds.

Meanwhile, tests on the transformed SCADI dataset found that the average accuracy value using the CBFS was higher than the average accuracy value of the two methods, which is 85.86%, with 16 features selected. Meanwhile, the average time required for Correlation-Base Feature Selection, 0.05 and the proposed threshold required the same execution time of 0.02 seconds.

5. Discussion

Based on the results of trials conducted using K-Cross Fold Validation with a value of K = 10 with 10 random seeds in ten original datasets and ten datasets that have been transformed using FFT and IFFT, it was found that:

5.1 Average accuracy value

A. Original dataset

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

The average accuracy value on trials using the original dataset showed that the proposed 60% threshold method produced a higher accuracy value than the CBFS algorithm, and 10% had the same average accuracy value.

2. The proposed threshold was compared with the threshold of 0.05.

   The average accuracy value on trials using the original dataset showed that the proposed 50% threshold method produced a higher accuracy value than the 0.05 threshold, and 30% had the same average accuracy value.

3. Comparison of the proposed threshold, 0.05 threshold and Correlation-Base Feature Selection.

   The average accuracy value of these three methods showed that the proposed 40% threshold results in a higher accuracy value of the two methods being compared. In addition, there was a 30% average accuracy value that had the same value. Broadly speaking, it can be concluded that the 70% average accuracy value using the proposed threshold had a better value.

B. Transformation Dataset

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

   The average accuracy value on trials using a transformed dataset shows that the proposed 50% threshold method produces a higher average accuracy value than the CBFS algorithm. Ten percent of it had the same average accuracy value as the Correlation-Base Feature Selection

2. The proposed threshold was compared to the threshold of 0.05.

   The average accuracy value on trials using the original dataset shows that the proposed threshold method of 40% produces an accuracy value that is higher than the threshold of 0.05, and 20% has the same average accuracy value.

3. Comparison of the proposed threshold, 0.05 threshold and Correlation-Base Feature Selection. The average accuracy value of these three methods shows that the proposed 20% threshold results in a higher accuracy value than the two methods being compared, and the 10% average accuracy value has the same value. Broadly speaking, it can be concluded that the 50% average accuracy value using the proposed threshold on the transformed data had a less good value than the 0.05 threshold and the Correlation-Base Feature Selection.

5.2 Average time required

A. Original Dataset

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

   By using the original dataset, the implementation of the proposed threshold in IG based feature selection resulted in less average execution time compared to using the CBFS algorithm. Among the ten datasets tested, it was found that at the proposed threshold of 50% of the dataset, it required an average execution time that was less than using the Correlation-Base Feature Selection, and 10% of the dataset required the same average time.

2. The proposed threshold was compared with the threshold of 0.05.

   Among the ten datasets tested, 70% of the dataset took less time than the use of a threshold of 0.05 and 20% of the dataset which required the same execution time.

3. Comparison of the proposed threshold, 0.05 threshold and Correlation-Base Feature Selection.

   By using the proposed threshold, it was found that 50% of the tested dataset required less time average than the 0.05 threshold and the Correlation-Base Feature Selection. IN addition, there were 10% of these datasets required the same average execution time, which is 0.25 seconds for the Contraceptive Method dataset.

B. Transformation Dataset

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

   By using the proposed threshold it is found that only 20% of the dataset produces less average time, and 30% has the same average time.

2. The proposed threshold was compared to the threshold of 0.05.

   At the proposed threshold, it was found that 40% of the transformed dataset had a faster average execution time than using a threshold of 0.05. In addition there were 20% yield the same time average.

3. The comparison of the proposed threshold, 0.05 threshold and Correlation-Base Feature Selection.

   Only 10% of the dataset produced a less average time than the 0.05 threshold and Correlation-Base Feature Selection, and 20% of the datasets had the same average time.

The average accuracy value and the average time required for the entire dataset used in this trial can be seen in Table 1, Table 2, Figure 3, Figure 4, Figure 5 and Figure 6.

Table 1. Test Results for Accuracy and Time Required

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	77,01%	4,17	2	63,16%	5,31	10	90,15%	4,91
Cancer	569	31	12	95,79%	0,08	26	96,63%	0,09	15	94,39%	0,07
Contraceptive Method	1.473	9	3	48,74%	0,25	3	48,74%	0,25	4	51,64%	0,25
Dermatology	366	33	15	94,92%	0,07	33	97,05%	0,05	26	97,43%	0,04
Divorce	170	54	6	96,53%	0,03	54	97,65%	0,03	52	97,65%	0,02
Electrical Grid	10.000	14	9	100,00%	0,91	5	100,00%	0,52	1	100,00%	0,17
CNAE-9	1.080	857	28	81,18%	0,27	57	87,56%	0,51	65	88,05%	0,58
Urban Land Cover	168	148	148	87,68%	0,06	110	85,71%	0,09	57	84,76%	0,07
Epilepsy	11.500	179	178	69,48%	15,72	178	69,60%	26,91	178	69,60%	26,91
SCADI	70	206	19	84,14%	0,02	123	83,43%	0,03	64	83,86%	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

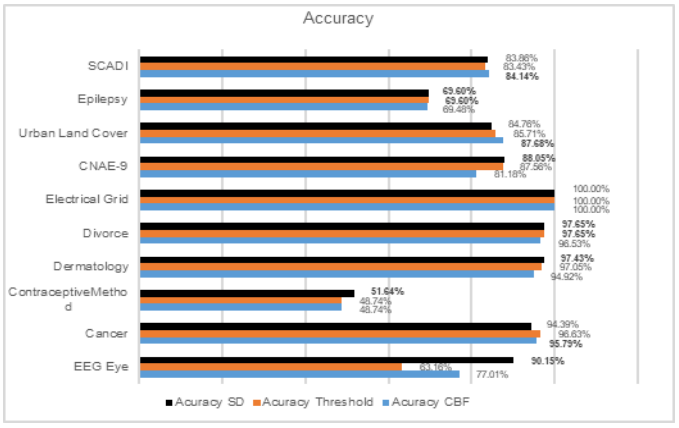

Fig.3. Comparison of average accuracy values

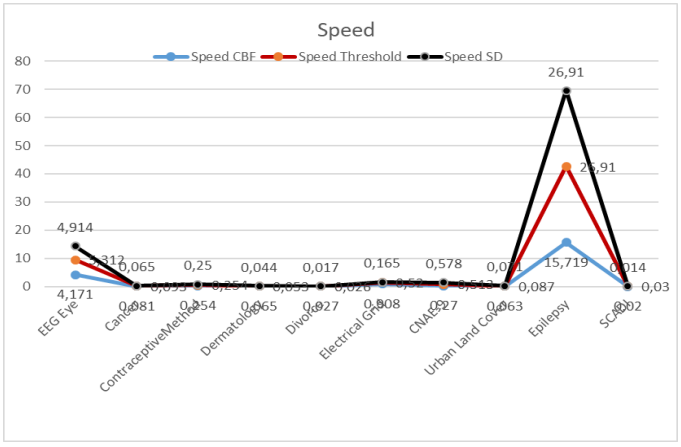

Fig.4. Comparison of the average required time

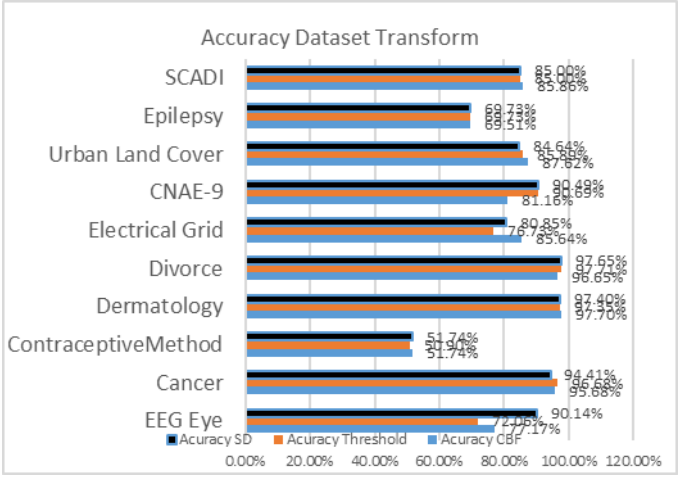

Fig 5. Comparison for the Mean Value of Transformation Dataset Accuracy

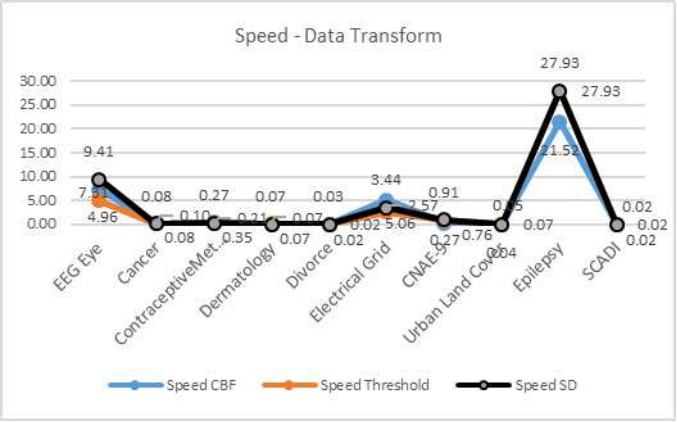

Fig.6. Comparison of the Average Time Required for the Transformation Dataset

6. Conclusion

From the trials that have been carried out on the original dataset, it can be concluded that

1. The average accuracy value in the original dataset show that the proposed threshold method produces a high average accuracy value compared to using the CBFS and the threshold of 0.05. There are 40% of the original dataset which produces a high average value of accuracy, and 30% of the dataset has a high average value whose value is the same as the threshold method of 0.05. So that 70% of the tested dataset produces a higher average value compared to the two methods being compared.
2. By using the proposed threshold value, it is found that 50% of the dataset used results in less average execution time than using CBFS and a threshold of 0.05. In addition, 10% of the dataset yields the same time average with the 0.05 threshold method. So that there are 60% of the tested dataset resulting in a shorter average time (faster) than the CBFS and the threshold of 0.05.

Meanwhile, for datasets that have undergone transformation using FFT and IFFT, it is found that:

1. By using the proposed threshold, the average accuracy value is less good than the CBFS and the threshold of 0.05. This can be seen from the results obtained in the trial, 70% of the dataset produces an average accuracy value that is higher than the average accuracy value generated by the proposed threshold.
2. The proposed threshold also results in a less than good average time on the transformed test dataset. Seventy percent of the transformed dataset took longer. Only 30% of the dataset requires less time than the 0.05 threshold method and CBFS.
3. By using the proposed threshold, the average accuracy value of the transformed dataset has increased within the range of 0.02% to 2.44%. Meanwhile, the average time required has decreased within the range of 0.01

seconds to 4.50 seconds.

4. Random Forest execution without using feature selection on the transformed dataset results in increased average accuracy value between 0.01% and 2.61%. This happened in 70% of the tested datasets. As for the average time needed, 60% of the transformed dataset takes less time than the original dataset, and 10% requires the same time. The difference in time needed is between 0.01 and 3.05 seconds.

Based on the results of trials that have been carried out, there are several things that need to be considered, including:

1. Dataset transformations cannot be used on datasets if the data is incomplete (missing value). Preprocessing is needed so that the dataset is complete.
2. The implementation of FFT and IFFT in the dataset needs to be considered, especially in the IG method with the proposed length.
3. Execution time (speed) and accuracy value are variables which are inversely proportional. So that it requires a choice which is preferred.

ABBREVIATIONS

CBFS: Correlation-Base Feature Selection

FFT: Fast Fourier Transform

IFFT: Inverse Fast Fourier Transform

IG: Information Gain

DECLARATIONS

Acknowledgments

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:

1. UCI Machine Learning Repository (www.arsip.Ics.uci.edu/ml)

2. Kaggle Datasets (www.kaggle.com/datasets)

Competing interests

The author reports no potential conflict of interest.

References

1. Wei G, Zhao J, Feng Y, He A, Yu J. A novel hybrid feature selection method based on dynamic feature importance. Applied Soft Computing. 2020;93:106337.

2. Prasetiyowati MI, Maulidevi NU, Surendro K. The Speed and Accuracy Evaluation of Random Forest Performance by Selecting Features in the Transformation Data. IEEA 2020: Proceedings of the 2020 The 9th International Conference on Informatics, Environment, Energy and Applications. 2020;125–130.

3. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research. 2003;3:1157–1182.

4. Ma J, Gao X. A filter-based feature construction and feature selection approach for classification using Genetic Programming. Knowledge-Based Systems. 2020;196:105806.

5. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis. 2020;143:106839.

6. Thabtah F, Kamalov F, Hammoud S, Shahamiri SR. Least Loss: A simplified filter method for feature selection. Information Sciences. 2020;534:1–15.

7. Samami M, Akbari E, Abdar M, Plawiak P, Nematzadeh H, Basiri ME, et al. A mixed solution-based high agreement filtering method for class noise detection in binary classification. Physica A: Statistical Mechanics and its Applications. 2020;553:124219.

8. Das H, Naik B, Behera HS. A Jaya algorithm based wrapper method for optimal feature selection in supervised classification. Journal of King Saud University - Computer and Information Sciences. 2020;S1319157820303670.

9. 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.

10. Lu M. Embedded feature selection accounting for unknown data heterogeneity. Expert Systems with Applications. 2019;119:350–61.

11. Zhang P, Gao W. Feature selection considering Uncertainty Change Ratio of the class label. Applied Soft Computing. 2020;95:106537.

12. Elmaizi A, Nhaila H, Sarhrouni E, Hammouch A, Nacir C. A novel information gain based approach for classification and dimensionality reduction of hyperspectral images. Procedia Computer Science. 2019;148:126–34.

13. Jadhav S, He H, Jenkins K. Information gain directed genetic algorithm wrapper feature selection for credit rating. Applied Soft Computing. 2018;69:541–53.

14. Singer G, Anuar R, Ben-Gal I. A weighted information-gain measure for ordinal classification trees. Expert Systems with Applications. 2020;152:113375.

15. Demsˇar J, Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7. 2006;7:1–30.

16. Yang Z, Ye Q, Chen Q, Ma X, Fu L, Yang G, et al. Robust discriminant feature selection via joint L 2 , 1 - norm distance minimization and maximization. Knowledge-Based Systems. 2020;106090.

17. Tsai C-F, Sung Y-T. Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. Knowledge-Based Systems. 2020;203:106097.

18. Ghaderi H, Kabiri P. Fourier transform and correlation-based feature selection for fault detection of automobile engines. The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012) [Internet]. Shiraz, Fars, Iran: IEEE; 2012 [cited 2020 Sep 18]. p. 514–9. Available from: http://ieeexplore.ieee.org/document/6313801/

19. Herff C, Krusienski DJ. Extracting Features from Time Series. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science [Internet]. Cham: Springer International Publishing; 2019 [cited 2020 Sep 18]. p. 85–100. Available from: http://link.springer.com/10.1007/978-3-319-99713-1_7

20. 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. Clinical Epidemiology and Global Health. 2019;7:274–8.

21. Hosseini S, Roshani GH, Setayeshi S. Precise gamma based two-phase flow meter using frequency feature extraction and only one detector. Flow Measurement and Instrumentation. 2020;72:101693.

22. Pratama SF, Muda AK, Choo Y-H. Arbitrarily Substantial Number Representation for Complex Number. Journal of Telecommunication, Electronic and Computer Engineering. 2018;10:23–6.

23. Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. :9.

24. Dua, D., Graff, C. UCI Machine Learning Repository [Internet]. University of California, School of Information and Computer Science.; Available from: http://archive.ics.uci.edu/ml

25. Breast Cancer Wisconsin (Diagnostic) Data Set Predict whether the cancer is benign or malignant [Internet]. Available from: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

26. Yöntem MK, Ilhan T. Divorce Prediction Using Correlation Based Feature Selection and Artificial Neural Networks. Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi; 2019.

27. Johnson BA. High-resolution urban land-cover classification using a competitive multi-scale object-based approach. Remote Sensing Letters. 2013;4:131–40.

28. Johnson B, Xie Z. Classifying a high resolution image of an urban area using super-object information. ISPRS Journal of Photogrammetry and Remote Sensing. 2013;83:40–9.

29. 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;64:061907.

30. Zarchi MS, Fatemi Bushehri SMM, Dehghanizadeh M. SCADI: A standard dataset for self-care problems classification of children with physical and motor disability. International Journal of Medical Informatics. 2018;114:81–7.

31. Fatemi Bushehri SMM, Zarchi MS. An expert model for self-care problems classification using probabilistic neural network and feature selection approach. Applied Soft Computing. 2019;82:105545.

32. Lei S. A Feature Selection Method Based on Information Gain and Genetic Algorithm. 2012 International Conference on Computer Science and Electronics Engineering [Internet]. Hangzhou, Zhejiang, China: IEEE; 2012 [cited 2020 Jul 18]. p. 355–8. Available from: http://ieeexplore.ieee.org/document/6188038/

33. Chen M-Y, Chen B-T. Online fuzzy time series analysis based on entropy discretization and a Fast Fourier Transform. Applied Soft Computing. 2014;14:156–66.

Figures

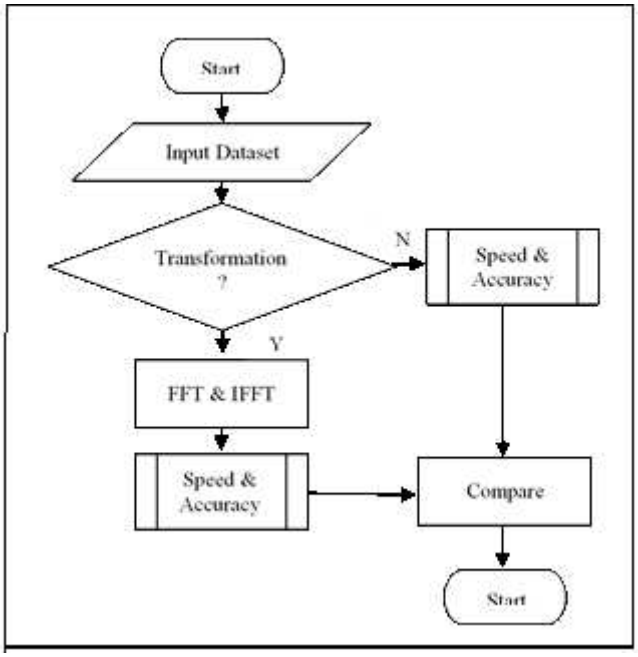

Figure 1

Research Stages

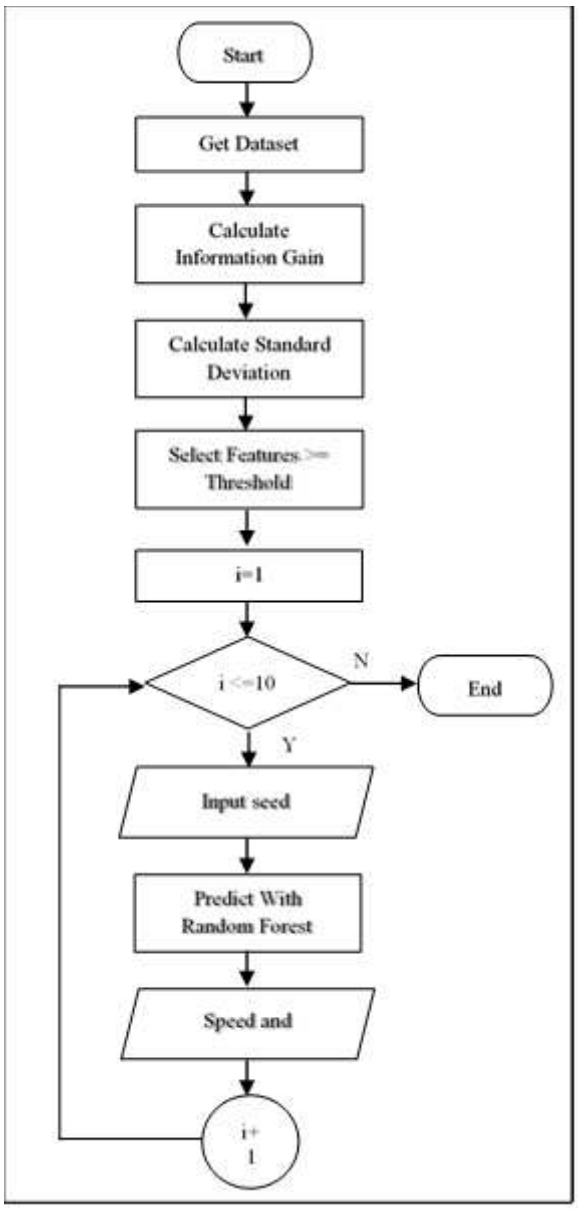

Figure 2

Fowchart Speed and Accuracy

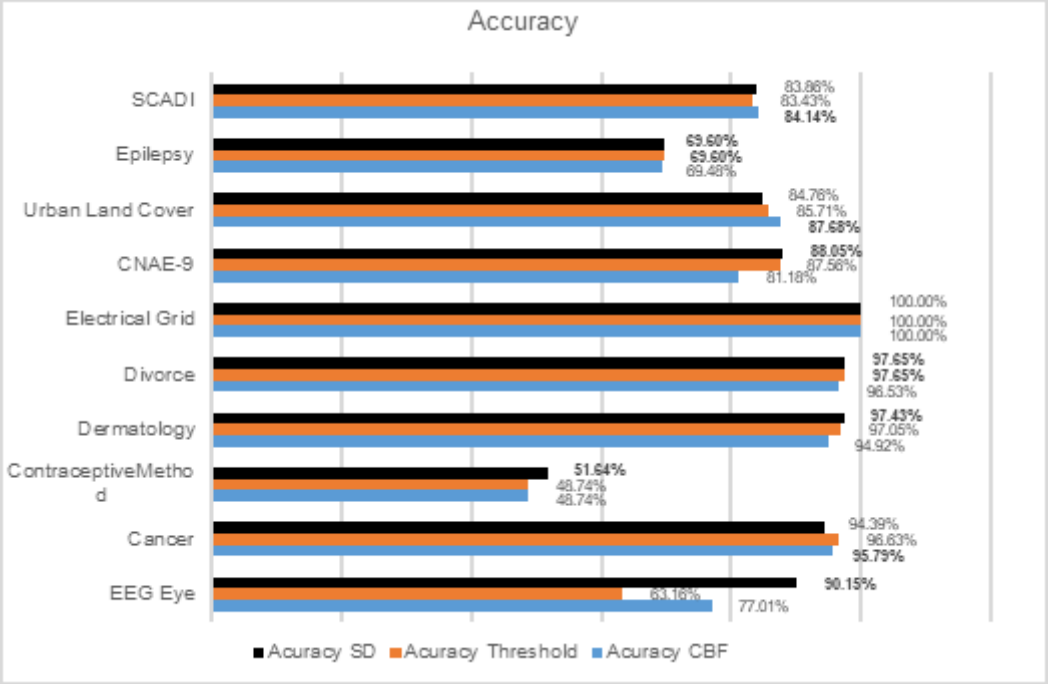

Figure 3

Comparison of average accuracy values

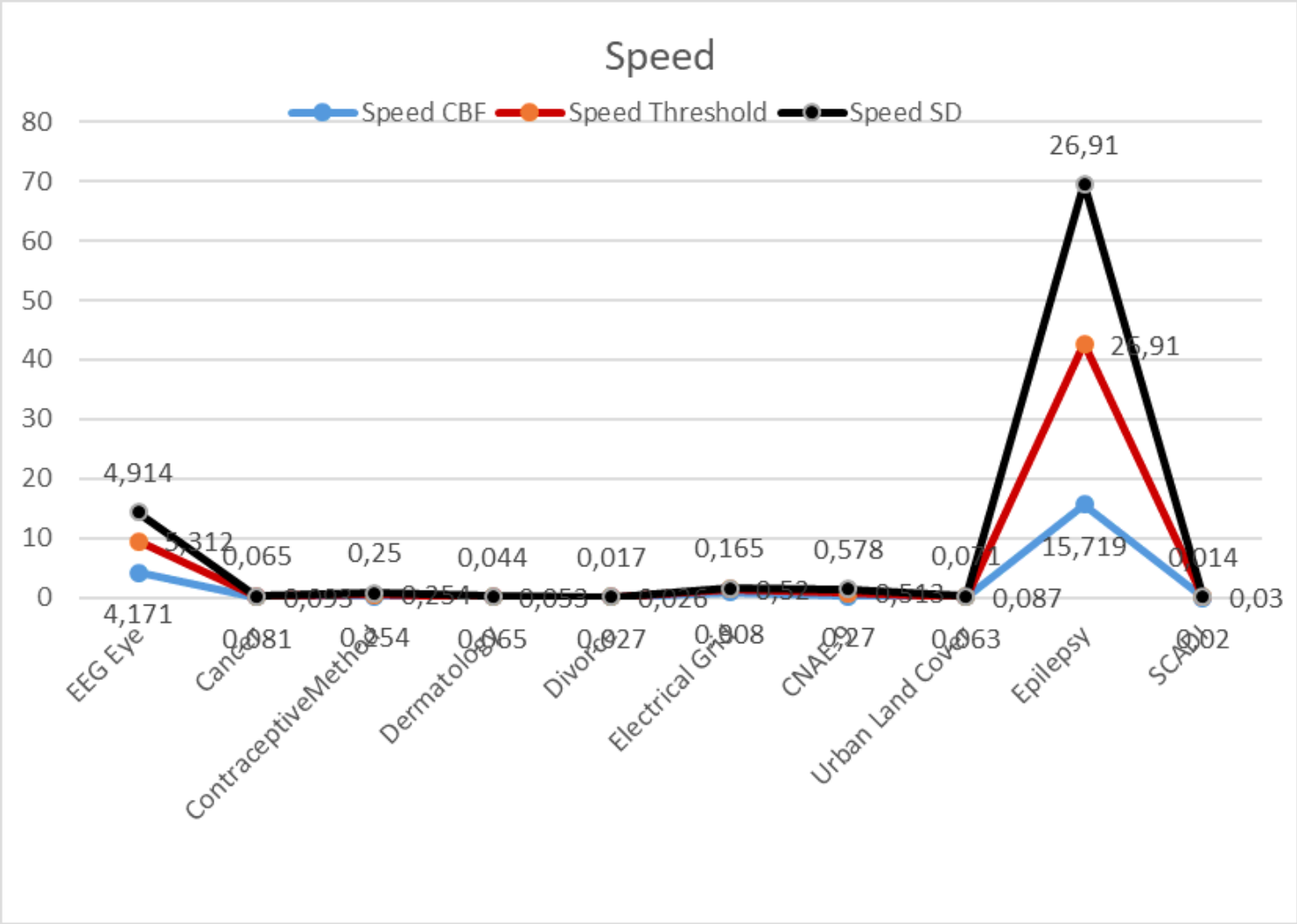

Figure 4

Comparison of the average required time

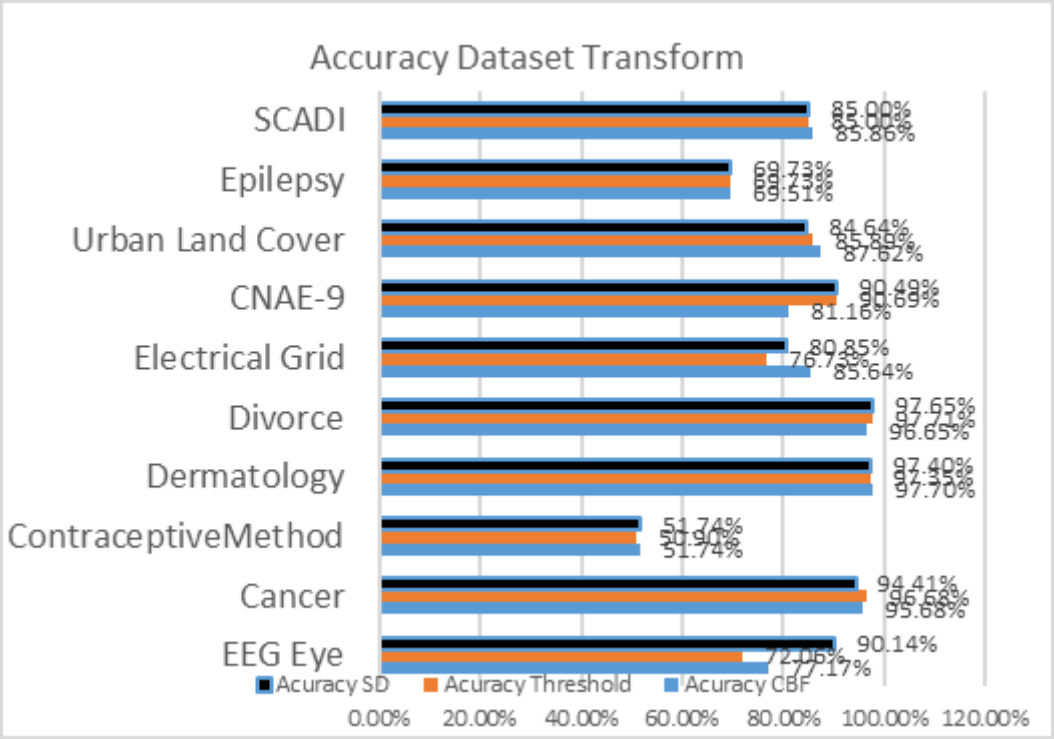

Figure 5

Comparison for the Mean Value of Transformation Dataset Accuracy

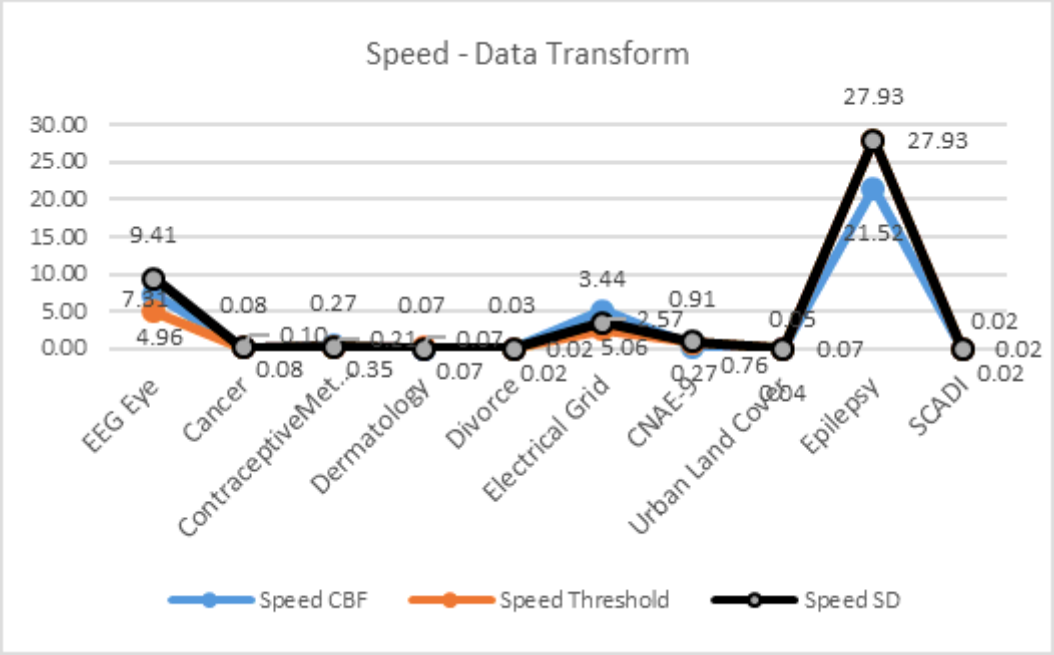

Figure 6

Comparison of the Average Time Required for the Transformation Dataset