Data Mining Based Artificial Intelligent Technique for Identifying Abnormalities from Brain Signal Data

Md. Nurul Ahad Tawhid[1](✉), Siuly Siuly[1], Kate Wang[2], and Hua Wang[1]

[1] Institute for Sustainable Industries and Liveable Cities, Victoria University, Melbourne, Australia
md.tawhid1@live.vu.edu.au, {siuly.siuly,hua.wang}@vu.edu.au

[2] School of Health and Biomedical Sciences, RMIT University, Melbourne, Australia
kate.wang@rmit.edu.au

Abstract. Analysis of brain signal data like Electroencephalography (EEG) plays an important role in efficient diagnosis of neurological disorders and treatment. EEG records electrical activity of the brain and contains huge volume of multi-channel time-series data that are visually analyzed by neurologists to identify abnormalities within the brain, which is time-consuming, error-prone, and subject to fatigue. Therefore, an automatic data mining system is always in need to detect abnormality from those large volume of data. To meet the requirements, in this study, a time-frequency spectrogram image-based classification framework is developed using texture feature extractor and machine learning (ML) based classifiers. At first, signals are filtered to remove noises and artifacts and normalized. Then signals are segmented into small chunks and spectrogram images are generated from those segments using short-time Fourier transform. After that, histogram based textural features are extracted and significant features are selected using principal component analysis. Finally, those features are fed into three ML based classifiers for categorizing the signals into different groups. The proposed system is tested on EEG brain signal data and have obtained promising results in identifying different abnormality groups, which indicates that the proposed system can be used for mining large volume of brain signal data.

Keywords: Brain signal data · EEG · Time-Frequency spectrogram image · Machine learning · Data mining

1 Introduction

In recent years, comprehensive studies have made on brain signal data, particularly using electroencephalogram (EEG) data due to its significant importance in health and medicine related applications [3]. Efficient and effective analysis of EEG signal is useful for various purposes like neurological diseases diagnosis and treatment [4,18,24,27], brain computer interface [20], sleep stage detection [21], emotion/fatigue detection [6,16] etc. EEG records the spontaneous electrical activity of the brain, which is large volume of time-series data. These data are

W. Zhang et al. (Eds.): WISE 2021, LNCS 13080, pp. 198–206, 2021.
https://doi.org/10.1007/978-3-030-90888-1_16

aperiodic, non-stationary and dynamic in nature and contains patterns related to the subject's mental health state [17]. Analysis of those large-scale aperiodic and non-stationary EEG signals is currently a challenging task. Data mining system can extract biomarkers from brain signal data and use those biomarkers to create computer aided diagnostic (CAD) systems that can classify brain states based on abnormalities.

Typically, EEG signal mining process can be divided into two steps: feature extraction and classification of the extracted features using different classifiers. Several techniques for analyzing and classifying large EEG signal data have been developed in recent years [11,17,19,22,25]. Most of these studies have used different statistical information as features for the signal classification with different classifiers. These traditional methods are often not feasible in extracting significant and discriminative features from large EEG data. Moreover, statistical analysis of larger recording data may overlook short-term changes in signal characteristics, which are important for abnormality detection. Using visual representation of short-term signal segments can solve this issue as it generates visual representation of raw data and works on small segments. Furthermore, most of the studies in this field have only tested their approaches on a single dataset, therefore their application to other datasets is debatable. Nonetheless, most investigations have focused on identifying one neurological disorder from EEG data (2-class problem). Few research have looked at identifying two neurological disorders from healthy control (HC) subjects (3-class) in the same system as, authors in [10,13] worked in detection of mild cognitive impairment and Alzheimer's disease patients from HC subjects. Authors of [2,9] have worked in classifying autism spectrum disorder (ASD) and epilepsy (EP) from HCs. But to the best of our knowledge, no research has conducted to detect more than two disorders in a single framework from HC subjects. This is because EEG data volume are huge in nature and the data has overlapping biomarkers for various diseases.

Therefore, to perform classification on these kind of overlapping feature-based data into multi-class requires special data mining techniques. Furthermore, a generic mining framework to conduct classification tasks and identify different types of abnormalities from EEG signals is required. This study aims to fill this gap by developing a data mining framework using short-term visual representation of the brain signal data to classify into multiple abnormality classes based on the biomarker presented in the visual representation of the signal data.

To achieve the aforementioned goal, we have developed a data mining framework for brain signal data, specifically for EEG, to identify four neurological abnormalities namely, ASD, EP, Parkinson's disease (PD), and Schizophrenia (SZ) from HC subjects (5 class) using time-frequency (T-F) spectrogram image and ML based classifiers. The brain signal data is initially filtered to remove noise and artifacts. The signals are then divided into small time frame windows, and spectrogram plotting images are created using a short-time fourier transform (STFT). Completed CENTRIST (cCENTRIST), a histogram-based feature extraction technique is used to extract textural information from those

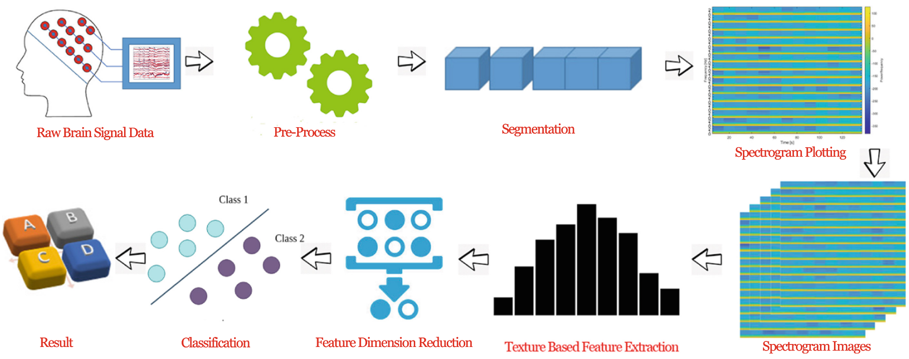


Fig. 1. Overview of the proposed framework.

images. The dimension of the retrieved features is then reduced using principal component analysis (PCA). Finally, three ML based classifiers are used to categorize the reduced extracted features: support vector machine (SVM), k-nearest neighbor (kNN) and random forest (RF).

This paper is organized as follows: Sect. 2 presents details workflow of the system. Section 3 states about used datasets and evaluation parameters. The experimental results are given in Sect. 4. Finally, Sect. 5 presents conclusion.

2 Workflow

In this study, we have used T-F based spectrogram image to classify brain signal data using cCENTRIST based feature extraction technique with three different ML based classification approaches: kNN, SVM and RF. The process consists of several steps: pre-processing, segmentation and spectrogram image generation, feature extraction and dimension reduction, and classification. Figure 1 depicts an overview of the proposed method. Details of those steps are given below.

2.1 Pre-processing the Brain Signal Data

In this step, we have pre-processed the brain signal data for removing the noise and artifacts introduced by the recording environment and the muscle movement of the subject during the recording time. These filtering processes are done due to some noise and artifacts are very much similar to some disease related signal patterns and may mislead the diagnosis process [25]. To perform the filtering, at first, we used the common average referencing (CAR) technique to remove the common noise and signals from all channels by removing the average signal from all electrodes. After that, artifacts introduced by muscle activity, eye movement and external noise are removed by passing the signal into a low pass infinite impulse response (IIR) filter with a cutoff frequency 40 Hz. Finally, the signals are normalized to a distribution of zero mean and a variance of one to reduce the individual signal differences and to reduce the computational complexity.

2.2 Spectrogram Image Generation

Here, the pre-processed signal data are converted into spectrogram images in two steps: at first the brain signal data are segmented into small chunks of three seconds (3 s) to increase the dataset size and as well as extract maximum number of features from the small signal segments. In this segmentation process original signals are segmented into small data chunks and given the level of original data, which makes an increase in the sample size. Then, spectrogram images are generated from those small chunks using STFT based plotting technique. These images provide a visual depiction of the signal data where different color represents the power and amplitude variation of the signal.

2.3 Feature Extraction and Dimension Reduction

Features from spectrogram images are extracted using cCENTRIST, an image feature extraction technique developed by Dey *et al.* [7] that combines Completed Local Binary Pattern (CLBP) and CENsus TRanform hISTogram (CENTRIST). It was developed by replacing Linear Binary Pattern (LBP) of CENTRIST [26] with CLBP and performed well on garments texture classification [7] and gender categorization from facial image [23]. cCENTRIST breaks the images into pyramid structure blocks and CLBP-based 3D histograms of those blocks are created and concatenated to produce a special histogram as an image feature. The retrieved feature's dimensions are reduced using PCA, and finally, the reduced feature set is utilized as input to various ML based classifiers.

2.4 Classification of the Extracted Features

To classify the cCENTRIST based histogram data of the spectrogram images, we have used three different ML based classifiers: RF, kNN and SVM. In kNN based classification, we have tested for 10 different k values (1 to 10) and for SVM, we used the same LibSVM [5] as the authors of cCENTRIST [7] used. Finally, these classifiers perform a multi-class classification for different neurological disorders and their performance are evaluated using different evaluation techniques.

3 Performance Evaluation Materials and Parameters

To validate the proposed model, we have used EEG brain signal data from four different neurological disorders: ASD, EP, PD and SZ. We have used these four datasets to perform a five-class classification using the proposed method (ASD vs EP vs PD vs SZ vs HC). Performance of the proposed method is evaluated using different evaluation matrices that are popular in this field of study.

3.1 Datasets

We have used four publicly available datasets of four different neurological abnormalities (ASD, EP, PD, SZ) to validate the proposed brain signal mining system. A brief information of those datasets are given in Table 1. Detail description of those datasets can be found in [1,4,12,14].

Table 1. Brief information of the used datasets.

Datasets	No of patients	No of HCs	Recording frequency	No of channels
ASD [1]	12	4	256	16
EP [14]	7	7	256	20
PD [4]	14	14	500	64
SZ [12]	14	14	250	19

3.2 Classification Performance Measure

This model is validated using a 5-fold cross-validation technique in order to reduce its classification bias and predict its overall accuracy across the entire dataset. This process involves dividing the dataset into five parts; four of which are used to train the classifier and the remainder one to test the learned system. This step is done five times so that each image in the dataset belongs to the test set exactly once. Finally, the 5-fold results are used to evaluate the system's performance using five parameters: sensitivity (Sen), specificity (Spec), precision (Prec), F1 score (F1), and accuracy (Acc). These criteria allow to predict the behavior of the classifiers on the test data [8,15,28].

4 Experimental Results

In this study, we have developed a brain signal data mining framework using spectrogram images of the signal data and ML based approaches. The proposed framework has tested on four neurological disease related EEG datasets and performed a five-class classification task. This section describes and visualizes the obtained results in detail with experimental setups.

4.1 Experimental Setup

Since EEG recordings contain varying sample rates and channels, the datasets must be standardized to make the data comparable. For this, we kept the ASD dataset (has 16 channel) as base and converted PD, EP, SZ datasets into that format by keeping data from standard 16 channels (Fp1, Fp2, F3, F4, F7, F8, C3, C4, T3, T4, P3, P4, T5, T6, O1, O2) and discarding other channel data and finally, resampled those 256 Hz. After that, EEG signals are pre-processed, segmented and spectrogram images are generated using STFT. This resulted in a total of 19417 images from four datasets, with ASD, EP, PD and SZ contributing 5437 (3825 ASD, 1612 HC), 2483 (1248 EP, 1235 HC), 1745 (864 PD, 881 HC) and 9752 (5312 SZ, 4440 HC) images, respectively. We combined all HC images to create a class of 8168 HC subjects, resulting in a 5-class classification problem.

4.2 Results

In this brain signal mining framework, we have used histogram-based cCENTRIST method to extract textural features from spectrogram images. The

Table 2. Five round average Sen, Spec, Prec, F1 and Acc for SVM, kNN and RF.

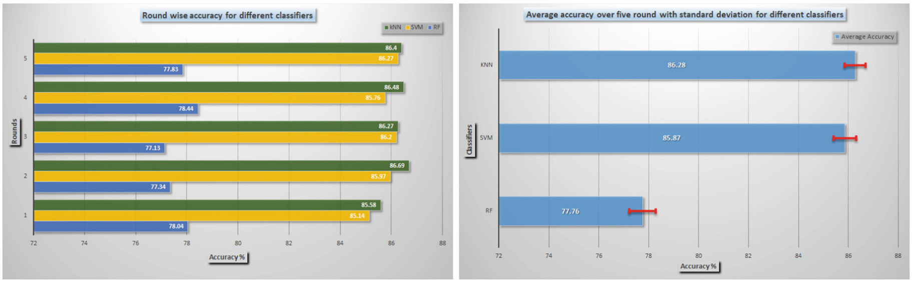
Disease	SVM				kNN				RF			
	Sen	Spec	Prec	F1	Sen	Spec	Prec	F1	Sen	Spec	Prec	F1
ASD	90.72	96.80	87.45	0.89	87.59	97.59	90.03	0.89	75.77	97.40	87.69	0.81
EP	83.59	98.51	79.46	0.81	77.14	98.06	73.29	0.75	31.87	99.97	98.56	0.48
Normal	84.77	87.83	83.49	0.84	88.89	86.56	82.77	0.86	92.29	70.17	69.20	0.79
PD	83.99	99.61	91.01	0.87	69.11	99.95	98.37	0.81	24.26	100.00	100.00	0.39
SZ	84.91	96.22	89.43	0.87	86.27	97.06	91.72	0.89	76.35	96.09	88.02	0.82
Avg	**85.59**	**95.79**	**86.17**	**0.86**	**81.80**	**95.85**	**87.23**	**0.84**	**60.11**	**92.72**	**88.69**	**0.66**
Acc	**85.87 ± 0.45**				**86.28 ± 0.42**				**77.76 ± 0.53**			

extracted features are then reduced in dimension using PCA, and classified using three ML-based classifiers: SVM, RF, and kNN (k = 1 to 10). Table 2 shows the 5-round average results of three classifiers for 5-fold cross validation technique. For kNN, we have just given the results of k=9 as it was the best of the ten different k values we tested.

Table 2 shows that kNN has the highest overall accuracy of 86.28%, whereas RF has the lowest overall accuracy of 77.76%. The accuracy of the SVM classifier is similar to that of the kNN. We have compared the accuracy of the three classifiers in Fig. 2, where Fig. 2a illustrates the round-wise accuracy and Fig. 2b depicts the average accuracy with standard deviation (SD).

For further evaluation of the proposed model, we have calculated and plotted the sensitivity, specificity, precision and F1 score for the classifiers, as shown in Fig. 3a, 3b, 3c and 3d.

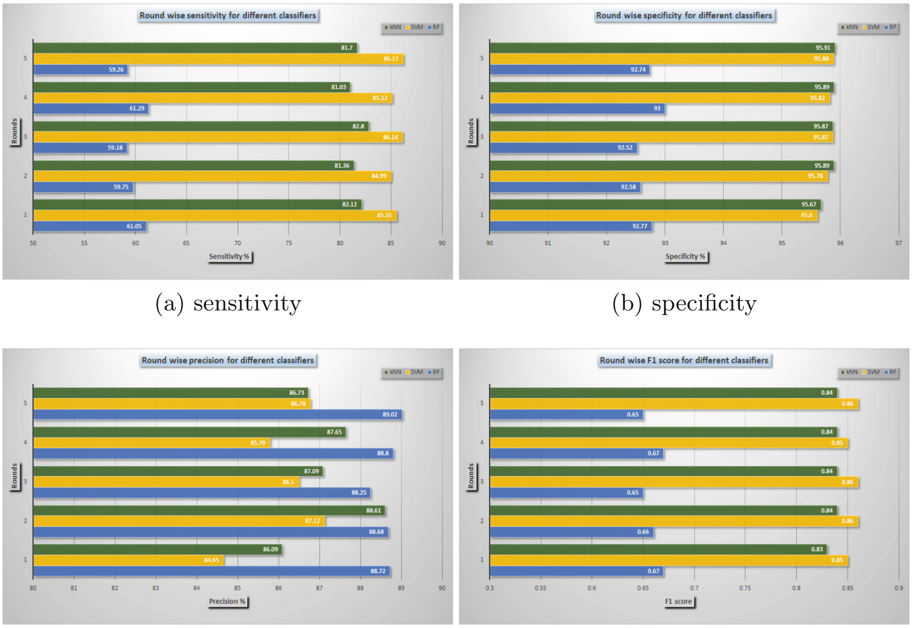
Figure 3a shows that, although kNN has the best classification performance, SVM clearly outperforms all other classifiers in terms of sensitivity, with a much higher round wise sensitivity value. SVM has the highest single round sensitivity value of 86.17% and an overall 5-fold average value of 85.59%(±0.55). RF



(a) Round wise accuracy

(b) Average accuracy with SD

Fig. 2. Accuracy comparison for different classifiers.



(a) sensitivity

(b) specificity

(c) precision

(d) F1 score

Fig. 3. Round wise Sen, Spec, Prec and F1 score comparison for different classifiers

has the lowest single round sensitivity of 59.18% with 5-fold average value of 60.11%(±1.00). For *k*NN, those values are 82.80% and 81.80% (±0.69), respectively. This result indicates that the SVM classifier is highly sensitive in detecting diseases than other classifiers.

Figure 3b shows the round-wise specificity of the used ML-based classifiers, showing that SVM and *k*NN have similar specificity values across rounds. The highest specificity values produced by *k*NN are 95.91% for a single round and 95.85%(±0.01) for a 5-round average. For SVM and RF, those values are 95.88%, 95.79 (±0.11), and 93.00%, 92.72% (±0.19), respectively. Higher specificity value indicates the model's ability to differentiate the healthy subjects from patients.

Round wise precision values of the three classifiers are plotted in Fig. 3c. Despite its poor overall performance, the RF classifier has a high precision for all rounds compared to other classifiers. This is because, despite its low sensitivity, the photos it identifies as the patient's image are true in general compared to other classifiers. Overall five round average precision for RF, *k*NN, SVM are 88.69% (±0.28), 87.23% (±0.96) and 86.17% (±0.98), respectively.

F1 score is the harmonic mean of precision and recall and an useful metric for evaluating classifier performance. Figure 3d depicts the round wise F1 score for the used classifiers, where SVM classifier outperforms other classifiers in all round values. Overall SVM has an average F1 score of 0.86 (±0.005) while for *k*NN, it is 0.84 (±0.005), and RF has the lowest average of 0.66 (±0.01).

5 Conclusion

In this study, a T-F spectrogram image with ML based data mining technique for brain signal data is proposed. To evaluate the proposed method, we have used EEG brain signal data for multi-category neurological diseases classification. The signal data is initially filtered to remove noise and artifacts, and segmented into small chunks. Then T-F based spectrogram images are generated from those segments using STFT. Textural features are extracted using cCENTRIST and PCA is used to reduce the extracted features dimension. Finally, kNN, SVM and RF classifiers are used for classifying those features into five classes: ASD, EP, PD, SZ, HC. Among the tested classifiers, kNN achieved the highest accuracy of 86.28%. Deep learning-based models, like convolutional neural networks, can be used in the future to classify the generated T-F based spectrogram images for mining brain signal data and improve classification performance.

References

1. Alhaddad, M.J., et al.: Diagnosis autism by fisher linear discriminant analysis FLDA via EEG. Int. J. Bio-Sci. Bio-Technol. **4**(2), 45–54 (2012)
2. Alturki, F.A., AlSharabi, K., Abdurraqeeb, A.M., Aljalal, M.: EEG signal analysis for diagnosing neurological disorders using discrete wavelet transform and intelligent techniques. Sensors **20**(9), 2505 (2020)
3. Alvi, A.M., Siuly, S., Wang, H.: Neurological abnormality detection from electroencephalography data: a review. Artif. Intell. Rev., 1–38 (2021). https://doi.org/10.1007/s10462-021-10062-8
4. Anjum, M.F., Dasgupta, S., Mudumbai, R., Singh, A., Cavanagh, J.F., Narayanan, N.S.: Linear predictive coding distinguishes spectral EEG features of Parkinson's disease. Parkinsonism Relat. Disord. **79**, 79–85 (2020)
5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 1–27 (2011)
6. Demir, F., Sobahi, N., Siuly, S., Sengur, A.: Exploring deep learning features for automatic classification of human emotion using EEG rhythms. IEEE Sens. J. **21**(13), 14923–14930 (2021)
7. Dey, E.K., Tawhid, M., Ahad, N., Shoyaib, M.: An automated system for garment texture design class identification. Computers **4**(3), 265–282 (2015)
8. He, J., Rong, J., Sun, L., Wang, H., Zhang, Y., Ma, J.: A framework for cardiac arrhythmia detection from IoT-based ECGS. World Wide Web **23**(5), 2835–2850 (2020)
9. Ibrahim, S., Djemal, R., Alsuwailem, A.: Electroencephalography (EEG) signal processing for epilepsy and autism spectrum disorder diagnosis. Biocybernetics Biomed. Eng. **38**(1), 16–26 (2018)
10. Ieracitano, C., Mammone, N., Hussain, A., Morabito, F.C.: A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia. Neural Netw. **123**, 176–190 (2020)
11. Oh, S.L., et al.: A deep learning approach for Parkinson's disease diagnosis from EEG signals. Neural Comput. Appl. **32**(15), 10927–10933 (2020)
12. Olejarczyk, E., Jernajczyk, W.: Graph-based analysis of brain connectivity in schizophrenia. PLoS One **12**(11), e0188629 (2017)

13. Oltu, B., Akşahin, M.F., Kibaroğlu, S.: A novel electroencephalography based approach for Alzheimer's disease and mild cognitive impairment detection. Biomed. Sig. Process. Control **63**, 102223 (2021)
14. Pereira, A., Fiel, J.: Resting-state interictal EEG recordings of refractory epilepsy patients (2019). https://doi.org/10.17632/6HX2SMC7NW.1
15. Sarki, R., Ahmed, K., Wang, H., Zhang, Y.: Automated detection of mild and multi-class diabetic eye diseases using deep learning. Health Inf. Sci. Syst. **8**(1), 1–9 (2020). https://doi.org/10.1007/s13755-020-00125-5
16. Şengür, D., Siuly, S.: Efficient approach for EEG-based emotion recognition. Electron. Lett. **56**(25), 1361–1364 (2020)
17. Siuly, S., Alcin, O.F., Bajaj, V., Sengur, A., Zhang, Y.: Exploring Hermite transformation in brain signal analysis for the detection of epileptic seizure. IET Sci. Meas. Technol. **13**(1), 35–41 (2018)
18. Siuly, S., et al.: A new framework for automatic detection of patients with mild cognitive impairment using resting-state EEG signals. IEEE Trans. Neural Syst. Rehabil. Eng. **28**(9), 1966–1976 (2020)
19. Siuly, S., Khare, S.K., Bajaj, V., Wang, H., Zhang, Y.: A computerized method for automatic detection of schizophrenia using EEG signals. IEEE Trans. Neural Syst. Rehabil. Eng. **28**(11), 2390–2400 (2020)
20. Siuly, S., Li, Y.: Discriminating the brain activities for brain-computer interface applications through the optimal allocation-based approach. Neural Comput. Appl. **26**(4), 799–811 (2015)
21. Supriya, S., Siuly, S., Wang, H., Zhang, Y.: EEG sleep stages analysis and classification based on weighed complex network features. IEEE Trans. Emerg. Top. Comput. Intell. **5**(2), 236–246 (2018)
22. Supriya, S., Siuly, S., Wang, H., Zhang, Y.: Automated epilepsy detection techniques from electroencephalogram signals: a review study. Health Inf. Sci. Syst. **8**(1), 1–15 (2020). https://doi.org/10.1007/s13755-020-00129-1
23. Tawhid, M.N.A., Dey, E.K.: A gender recognition system from facial image. Int. J. Comput. Appl. **180**(23), 5–14 (2018)
24. Tawhid, M.N.A., Siuly, S., Wang, H.: Diagnosis of autism spectrum disorder from EEG using a time-frequency spectrogram image-based approach. Electron. Lett. **56**(25), 1372–1375 (2020)
25. Tawhid, M.N.A., Siuly, S., Wang, H., Whittaker, F., Wang, K., Zhang, Y.: A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from EEG. Plos One **16**(6), e0253094 (2021)
26. Wu, J., Rehg, J.M.: Centrist: a visual descriptor for scene categorization. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1489–1501 (2010)
27. Yin, J., Cao, J., Siuly, S., Wang, H.: An integrated mci detection framework based on spectral-temporal analysis. Int. J. Autom. Comput. **16**(6), 786–799 (2019)
28. Zhang, F., Wang, Y., Liu, S., Wang, H.: Decision-based evasion attacks on tree ensemble classifiers. World Wide Web **23**(5), 2957–2977 (2020). https://doi.org/10.1007/s11280-020-00813-y