Turkish Journal of Science & Technology
18(1), 215-222, 2023

Research Paper
https://doi.org/10.55525/tjst.1244759

BP19: An Accurate Audio Violence Detection Model Based On One-Dimensional Binary Pattern

Arif Metehan YILDIZ[1*], Tugce KELEŞ[2], Kubra YILDIRIM[3], Sengul DOGAN[4], Turker TUNCER[5]
[1,2,3,4,5] Department of Digital Forensics Engineering, College of Technology, Firat University, Elazig, Turkey
[*1] a.metehanyildiz@gmail.com, [2] tkeles@firat.edu.tr, [3] kubra.yildirim@firat.edu.tr, [4] sdogan@firat.edu.tr, [5] turkertuncer@firat.edu.tr

(Geliş/Received: 31/01/2023; Kabul/Accepted: 01/03/2023)

Abstract: Audio violence detection (AVD) is a hot-topic research area for sound forensics but there are limited AVD researches in the literature. Our primary objective is to contribute to sound forensics. Therefore, we collected a new audio dataset and proposed a binary pattern-based classification algorithm.
Materials and method: In the first stage, a new AVD dataset was collected. This dataset contains 301 sounds with two classes and these classes are violence and nonviolence. We have used this dataset as a test-bed. A feature engineering model has been presented in this research. One-dimensional binary pattern (BP) has been considered to extract features. Moreover, we have applied tunable q-factor wavelet transform (TQWT) to generate features at both frequency and space domains. In the feature selection phase, we have applied to iterative neighborhood component analysis (INCA) and the selected features have been classified by deploying the optimized support vector machine (SVM) classifier.
Results: Our model achieved 97.01% classification accuracy on the used dataset with 10-fold cross-validation.
Conclusions: The calculated results clearly demonstrated that feature engineering is the success solution for violence detection using audios.

Key words: Audio violence detection, feature engineering, machine learning, sound forensics

BP19: Tek Boyutlu İkili Modele Dayalı Doğru Bir Sesli Şiddet Tespit Modeli

Özet: Sesten şiddet tespiti (AVD), ses adli bilişimi için oldukça taze bir araştırma alanıdır, ancak literatürde sınırlı AVD araştırmaları vardır. Öncelikli hedefimiz ses adli bilişimine katkıda bulunmaktır. Bu nedenle, yeni bir ses veri seti topladık ve ikili örüntü tabanlı bir sınıflandırma algoritması önerdik.
Gereç ve yöntem: İlk aşamada yeni bir AVD veri seti toplandı. Bu veri seti iki sınıflı 301 ses içerir ve bu sınıflar şiddet ve şiddetsizliktir. Bu veri setini test ortamı olarak kullandık. Bu araştırmada bir özellik mühendisliği modeli sunulmuştur. Öznitelikleri çıkarmak için tek boyutlu ikili model (BP) düşünülmüştür. Ayrıca, hem frekans hem de uzay alanlarında özellikler oluşturmak için ayarlanabilir q-faktörü dalgacık dönüşümü (TQWT) uyguladık. Öznitelik seçimi aşamasında yinelemeli komşuluk bileşen analizi (INCA) uygulanmış ve seçilen öznitelikler optimize edilmiş destek vektör makinesi (SVM) sınıflandırıcı kullanılarak sınıflandırılmıştır.
Bulgular: Modelimiz, kullanılan veri setinde 10 kat çapraz doğrulama ile %97.01 sınıflandırma doğruluğu elde etti.
Sonuçlar: Hesaplanan sonuçlar, özellik mühendisliğinin ses kullanarak şiddet tespiti için başarılı bir çözüm olduğunu açıkça göstermiştir.

Anahtar kelimeler: Sesli şiddet tespiti, özellik mühendisliği, makine öğrenimi, ses adli bilişimi

1. Introduction

Speech is defined as the verbal transmission or expression of thoughts. On the other hand, communication is the exchange of thoughts, behaviors and information in which there is a sender-receiver relationship between any individual or group of individuals. When the relationship between these two concepts is examined, a multidimensional situation naturally emerges. First of all, making sense of how this relationship between individuals is established poses a difficult problem for machine learning methods. For example, with the combination of understanding a speech, communication patterns, the environment in which the communication is established and the surrounding context, a subjective situation that can naturally differ from individual to

[*] Corresponding author: a.metehanyildiz@gmail.com. ORCID Number of authors: [1] 0000-0003-0451-8600, [2] 0000-0003-0131-2826, [3] 0000-0002-4738-2777, [4]0000-0001-9677-5684, [5]0000-0002-1425-4664

individual arises. Because each individual is triggered differently by different factors in his own particular, and this situation is then called as different character structure. Therefore, there is a need to use many more models at the same time [1]. This environment becomes more complex when a different language or violent speech content is involved.

As a result, violent behaviors can harm other people in the environment and the individual. The emotional and instinctive states of the individuals who engage in these behaviors in the communication environment, together with the degree of violence, can determine whether or not there is any aggression. On the other hand, it examines the relationship between aggression and behavior and the relationships between subjective feelings and objective actions [2].

When various studies in the literature are examined, it is seen that some intimate-partner (lovely) relationships carry out a process that can cause serious injuries and even death, both psychologically and physically [3]. This environment, which takes place with the thought of keeping the partner under power and control in these relationships, creates a potential context for excessive emotions and conflicts [4]. On the other hand, some researchers have had difficulty recognizing emotions accurately due to personal emotions' ambiguity and multifaceted nature [5, 6]. Researchers have attempted to recognize and detect emotions using speech recognition [7, 8], violence analysis from images [9-11], and Natural Language Processing (NLP), which are generally applied to large datasets created using social media. In order to understand the emotion of interactions, applications such as Instagram, Facebook, Twitter, and Youtube were used [12-15].

NLP forms a fundamental link in the field of artificial intelligence and computer science. NLP studies include theory and methods that enable effective communication between humans and computers in natural language [16]. For these reasons, it is seen that it is frequently used to interpret, decode and make sense of human languages in a meaningful way. These features can help determine the relationship between individuals, conflict of emotions and level of violence. In recent years, many studies have been carried out to detect negative emotions such as abusive use of language and cyberbullying in online conversations. However, it is relatively difficult in real-life ritual to obtain a live recording of any setting [17, 18]. For example, when fully focused on the speech content, other details in the environment (intensity of sound, frequency of sound, physical characteristics of the environment, size of the environment, distance, etc.) are ignored. If we look at this event from the point of view of forensic information, it is necessary to evaluate the conversations and the environment in which the conversation takes place as a whole. In addition to these, one of the important points in determining the intensity is the entropy criterion. This process, which is based on sudden changes in the energy level of the audio signal, is used as an auxiliary argument in determining the intensity [19].

Collecting accurate and unmodified audio data is also a difficult process for forensic experts. Otherwise, obtaining erroneous results casts a shadow over a judicial review. Therefore, it is very important to be able to detect violence in terms of sound forensics. Based on the existing literature, it is seen that although audio is a very useful source of information, can be processed more simply than videos, and can be self-sufficient in terms of expressing violence most of the time, it is often overlooked.

A new sound classification model has been proposed using lightweight methods to detect intensity over sound, which is a difficult process in this research area. This model will be simply applied to solve sound classification problems with high success rate. With this proposed method, it is desired to solve the important problems in the literature.

1.1. Motivation and contributions

Sound forensics is one of the main branches of digital forensics. Therefore, it is a very important research area for gathering digital evidence but the incident response/digital forensics experts have extracted digital evidence manually and it is a time-consuming and difficult process[20]. Therefore, digital forensics professionals need an intelligent assistant to extract more digital evidence. Thus, we collected a new AVD dataset by using public sound sources. In order to get high classification performance on the collected audio dataset, we have recommended a feature engineering model by mimicking a deep learning structure since we have presented a multileveled feature extraction method. Furthermore, the presented model is a lightweight sound classification model since it uses lightweight methods to get classification results.

The contributions of our proposal are:

- A new AVD dataset was collected.
- A simple model has been presented and this model has got high classification performance.
- This model is an accurate model and can easily be applied to solve other sound classification problems

2. Dataset

We collected an AVD dataset from YouTube. This dataset has two classes and these classes are (i) violence and (ii) nonviolence. We only collected the sounds from the videos and we segmented these sounds. The length of the created each segment is about 2/3 seconds. The violence class has 154 sound observations and there are 147 sounds in the nonviolence class. The sample sounds have been demonstrated in Figure 1.

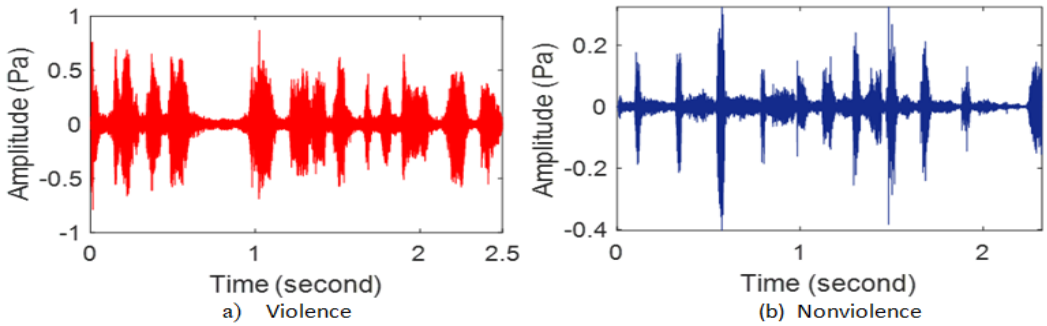

Figure 1. The sample sounds of the gathered dataset.

3. Our Proposal

The essential objective of the proposed model is to get high classification performance with a simple and lightweight feature engineering model. Our model uses TQWT and the one-dimensional BP to get features [21, 22]. The top of these features have been selected by deploying INCA [23] selector and the optimized SVM has been used to get the prediction vector. The schema of our proposal is demonstrated in Figure 2.

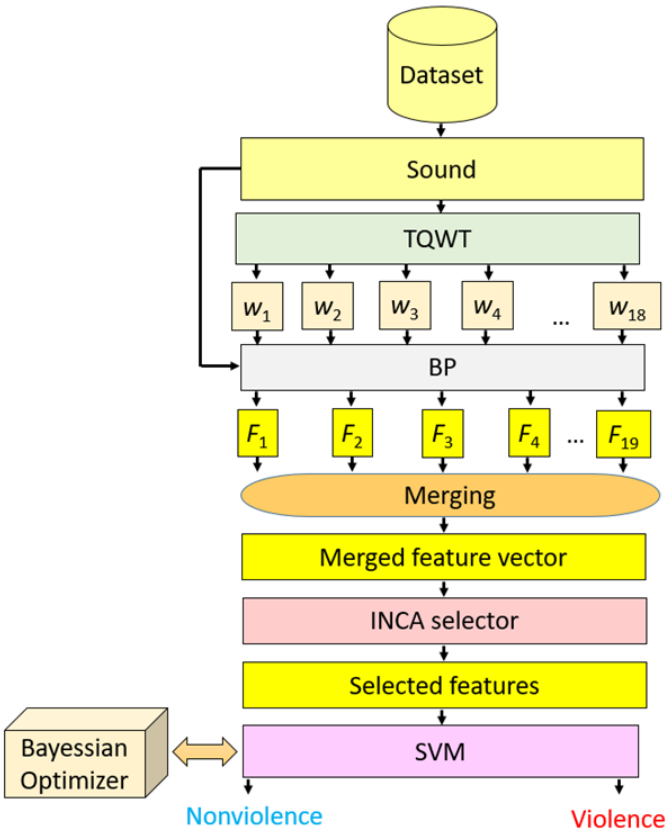

Figure 2. Schema of the proposed model. Herein, w: TQWT bands, F: feature vector.

The steps of our proposal are also given below for better explanation.
Step 0: Load the collected AVD dataset.
Step 1: Apply TQWT to the sound signal to generate 18 wavelet bands. Herein, the used parameters of the TQWT are 2,3 and 17. The given three parameters are named q-factor (q), redundancy value (r) and the number of levels (L) and TQWT-generated L+1 wavelet bands.
Step 2: Apply BP to the raw sound and the wavelet bands generated. The used feature extractor (BP) has been clarified below. In this step, 19 (by deploying 18 wavelet bands + 1 raw sound signal) feature vectors have been created and the length of each feature vector is equal to 256.
Step 2.1: Create overlapping blocks and the length of each block is nine.
Step 2.2: Generate binary features by deploying the center value and other values of each block.
Step 2.3: Transform/convert binary features to a decimal value and create a feature map signal.
Step 2.4: Extract the histogram of the feature map signal and obtain a feature vector with a length of 256.
The graphical denotation of the one-dimensional BP is depicted in Figure 3.

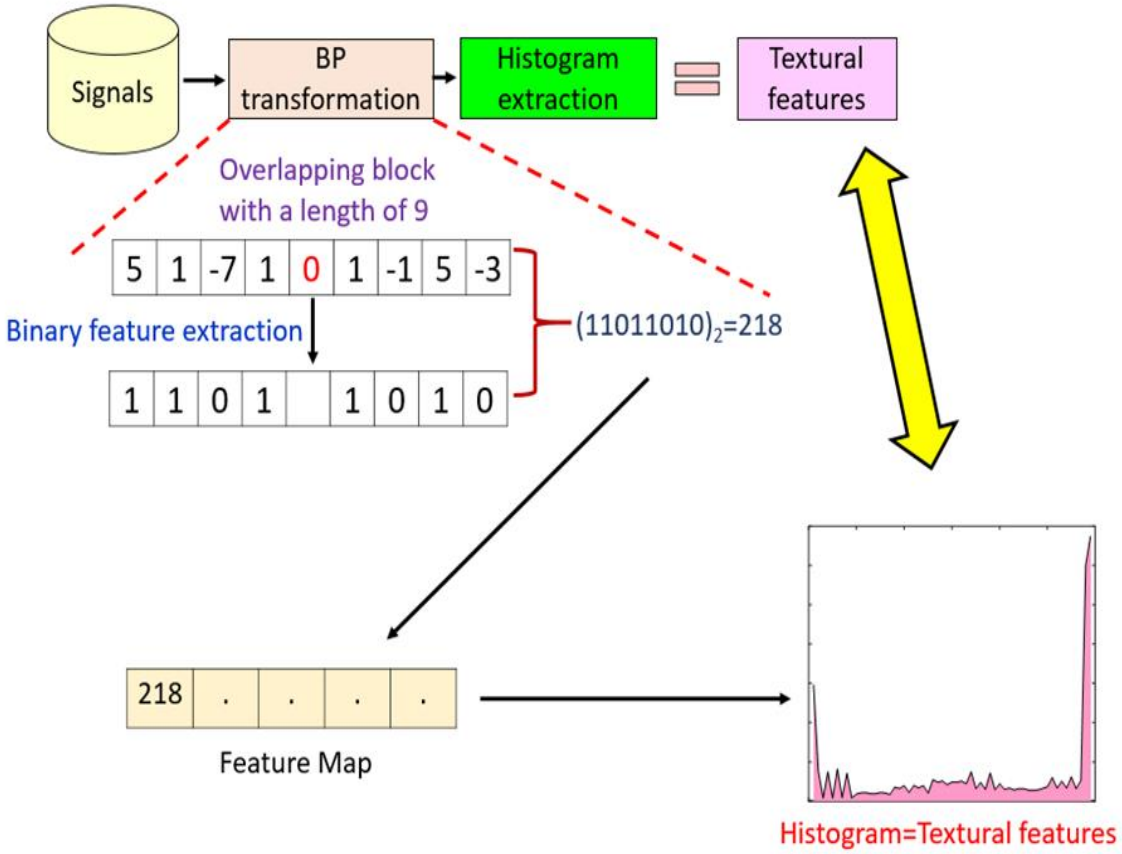

Figure 3. Graphical explanation of the BP feature extractor.

Step 3: Merge the generated 19 feature vectors and create the final feature vector with a length of 4864 (=256×19).
Step 4: Apply iterative INCA and choose the top 966 features from the created 4864 features.
Step 5: Classify these features by deploying the optimized SVM.
The given five steps above have been explained in our proposal.

4. Performance Evaluation

In this research, we have used a TQWT and BP-based model. BP generates features from 19 signals. Hence, this model is named BP19, like VGG19 or DarkNet19. However, this model is a lightweight model. We used MATLAB (2022a) programming environment to implement this model. We did not use any graphical processing

unit or parallel programming technique to get classification results. Furthermore, BP19 is a parametric classification model. The parameters used to create BP19 are listed in Table 1.

Table 1. The parameters used to create BP19

Method	Parameters	Output
TQWT [21]	q:2, r:3, J:17	18 wavelet bands
BP [22]	Length of overlapping block: 9, kernel: signum	256 features
Our feature extraction method	19 input (18 wavelet bands + 1 raw sound)	4864 features
INCA [23]	Loop range: [100-1000], Classifier: kNN	966 features
SVM [24]	Kernel: Gaussian, kernel scale: 86.94, Box constraint: 17.83, validation: 10-fold cross-validation	Results
Bayesian optimizer [25]	Fitness function: Maximizing accuracy, number of iterations: 1000	The optimized parameters of the SVM

The proposed BP19 has been created by deploying these parameters which are listed in Table 1. This problem is a binary classification problem [26]. Therefore, we have used accuracy, sensitivity, specificity and geometric mean performance metrics. In order to calculate these performance metrics, we have given the computed confusion matrix as below (see Figure 4).

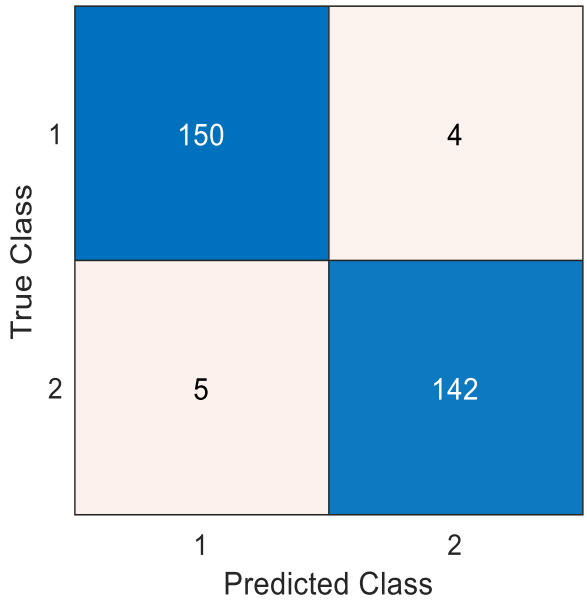

Figure 4. The calculated confusion matrix. 1: Violence, 2: Control.

The evaluation metrics are mathematically defined below[27].

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$Geometric\ Mean = \sqrt{(\frac{TP}{TP + FN})(\frac{TN}{TN + FP})} \quad (4)$$

The calculated performance values per Figure 4 are listed in Table 2.

Table 2. The calculated performance results (%) by deploying our proposed BP19.

Performance metric	**Result(s)**
Accuracy	97.01
Sensitivity	97.40
Specificity	96.60
Geometric mean	97

Table 2 demonstrated that our model reached over 96% classification performance per the used all performance metrics.

5. Ablation

In this section, our primary goal is to demonstrate the superiority of BP19. Therefore, we have applied BP to the audio signals, and the calculated results are given in this section. Therefore, we have created two cases and these cases are defined below.
Case 1: By applying BP for features extraction and the generated features are fed to the SVM classifier.
Case 2: By applying BP for features extraction and the top features have been selected by deploying INCA. The selected features are classified by deploying INCA.
The ablation results are shown in Figure 5.

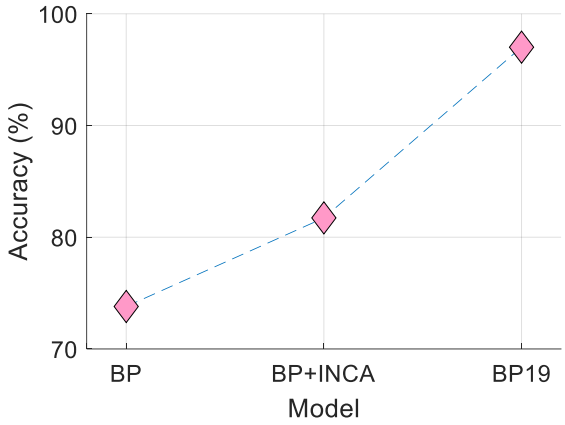

Figure 5. Accuracies of the models. BP: Case 1, BP+INCA: Case 2, BP19: Our model.

6. Feature Analysis

In the proposed BP19, we have used 19 signals. In order to detect the classification contribution of the generated 19 one-dimensional signals, we counted the selected features per the used signal and the feature analysis results have been depicted in Figure 6.

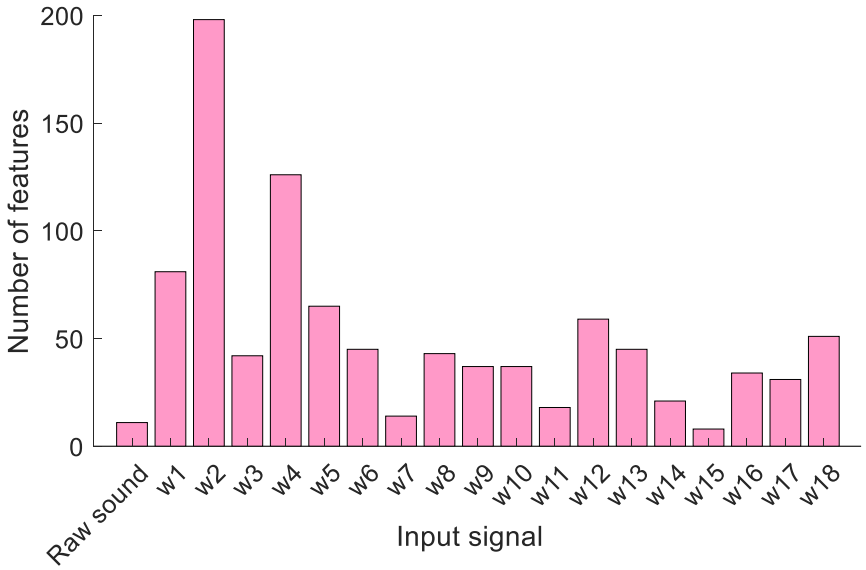

Figure 6. The number of selected features from the used input signals.

Figure 6 demonstrated that the 198 out of the generated 256 features from w2 (the generated second wavelet band by TQWT) are selected as top feature by INCA. Thus, the most valuable signal is w2. Only 11 features from the raw sound/audio signal. Per the number of feature, the worst band is w15 since INCA selected only 8 features from this signal.

7. Conclusions

We presented the BP19 model, and this model got results from the collected AVD dataset. Our collected dataset has 301 audios and these are categorized into (i) violence and (ii) control. The proposed BP19 creates 19 signals and generates features from these signals by deploying BP function. INCA selected the most informative 966 features, which have been classified by employing an optimized SVM with 10-fold cross-validation. Our model attained 97.01% classification accuracy and 97% geometric mean. Furthermore, we analyzed features per the generated signals and gave ablation results to show the superiority of our model. Finally, we investigated the classification ability of a feature engineering model. Results and findings demonstrate that high classification performance was attained with a simple model of the violence audio.

In the near future, we are planning to collect a big AVD dataset and propose self-organized feature engineering models. By using these models, automated AVD applications can be presented.

Acknowledgement

I would like to thanks all the authors who contributed to science by working in the field of Digital forensic while writing this article, and my advisor Turker Tuncer and head of our department Sengul Dogan who contributed to the creation of the article.

References

[1] Anwar A, Kanjo E, Anderez DO. DeepSafety: Multi-level Audio-Text Feature Extraction and Fusion Approach for Violence Detection in Conversations. arXiv e-prints 2022; arXiv:2206.11822.

[2] Baumeister RF, Bushman BJ. Emotions and aggressiveness, International handbook of violence research. Heitmeyer W, Hagan J, editors. Springer Dordrecht, 2003; 479–493.

[3] Allen T, Novak SA, Bench LL. Patterns of injuries: accident or abuse. Violence Against Wom 2007;13(8):802–16.

[4] Bulut M, Aslan R, Arslantaş H. Kabul Edilmemesi Gereken Toplumsal Bir Gerçek: Yakn Partner Şiddeti. Sakarya Tıp Dergisi 2020; 10(2):334–347.

[5] Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media; 15-18 May 2017; Montreal, Quebec, Canada. pp. 512–515.

[6] Bhavan A, Chauhan P, Shah RR, others. Bagged support vector machines for emotion recognition from speech. Knowl-Based Syst 2019; 184:104886.
[7] Atmaja BT, Akagi M. Speech emotion recognition based on speech segment using LSTM with attention model. In: 2019 IEEE International Conference on Signals and Systems (ICSigSys); 16 – 18 July 2019; Bandung, Indonesia. pp. 40–44.
[8] Li Y, Zhao T, Kawahara T, others. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In: Interspeech. 15-19 September 2019; Graz, Austria. pp. 2803–2807.
[9] Hajarolasvadi N, Demirel H. Deep facial emotion recognition in video using eigenframes. IET Image Process 2020; 14(14):3536–3546.
[10] Hu M, Wang H, Wang X, Yang J, Wang R. Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks. J Vis Commun Image R 2019; 59:176–185.
[11] Du Z, Wu S, Huang D, Li W, Wang Y. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. IEEE T Affect Comput 2019; 12(3):565–578.
[12] Plaza-del-Arco FM, Mart\'\in-Valdivia MT, Urena-Lopez LA, Mitkov R. Improved emotion recognition in Spanish social media through incorporation of lexical knowledge. Future Gener Comp Sy 2020; 110:1000–1008.
[13] Yang CT, Chen YL. Dacnn: Dynamic weighted attention with multi-channel convolutional neural network for emotion recognition. In: 2020 21st IEEE international conference on mobile data management (MDM); 30 June - 3 July 2020; Versailles, France. pp. 316–321.
[14] Batbaatar E, Li M, Ryu KH. Semantic-emotion neural network for emotion recognition from text. IEEE access 2019; 7:111866–111878.
[15] Abdullah NSD, Zolkepli IA. Sentiment analysis of online crowd input towards brand provocation in Facebook, Twitter, and Instagram. In: Proceedings of the International Conference on Big Data and Internet of Thing; 20 - 22 December 2017; London, United Kingdom. pp. 67–74.
[16] Kang Y, Cai Z, Tan CW, Huang Q, Liu H. Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics 2020; 7(2):139–172.
[17] Sharma HK, Kshitiz K, others. Nlp and machine learning techniques for detecting insulting comments on social networking platforms. In: 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE); 22-23 June 2018; Paris, France. pp. 265–272.
[18] Mossie Z, Wang JH. Vulnerable community identification using hate speech detection on social media. Inf Process Manag 2020; 57(3):102087.
[19] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, Violence content classification using audio features, in Advances in Artificial Intelligence: 4th Helenic Conference on AI, SETN 2006; 18-20 May 2006; Heraklion, Crete, Greece. pp. 502-507.
[20] Khanafseh M, Qatawneh M, Almobaideen W. A survey of various frameworks and solutions in all branches of digital forensics with a focus on cloud forensics. Int J Adv Comput Sci Appl 2019; 10(8).
[21] Subasi A, Tuncer T, Dogan S, Tanko D. Local binary pattern based feature extraction and machine learning for epileptic seizure prediction and detection. Modelling and Analysis of Active Biopotential Signals in Healthcare Volume 2 , Bristol, UK : IOP Publishing, 2020; pp. 6-1 to 6-31.
[22] Zeng W, Yuan J, Yuan C, Wang Q, Liu F, Wang Y. A new approach for the detection of abnormal heart sound signals using TQWT, VMD and neural networks. Artif Intell Rev 2021; 54(3):1613–47.
[23] Tuncer T, Ertam F. Neighborhood component analysis and reliefF based survival recognition methods for Hepatocellular carcinoma. Physica A: Statistical Mechanics and its Applications 2020; 540:123143.
[24] Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. A fast iterative nearest point algorithm for support vector machine classifier design. IEEE T Neural Networ 2000; 11(1):124–36.
[25] Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. Journal of Electronic Science and Technology. 2019; 17(1):26–40.
[26] Primus P, Haunschmid V, Praher P, Widmer G. Anomalous Sound Detection as a Simple Binary Classification Problem with Careful Selection of Proxy Outlier Examples. In:Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020); 2–3 November 2020; Tokyo, Japan. pp. 170-174.
[27] Dogan S, Barua PD, Kutlu H, Baygin M, Fujita H, Tuncer T, et al. Automated accurate fire detection system using ensemble pretrained residual network. Expert Syst Appl 2022; 203:117407.