Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 2, 2026, 197-207.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Ensemble Learning for Speech Emotion Recognition using Graph-Based Signal Dynamics

Zeynab Mohammadpoory [1,*], Mahda Nasrolahzadeh[2], and Sekineh Asadi Amiri [3]

[1] *Faculty of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran.*
[2] *Department of Electrical Engineering, University of Torbat Heydarieh, Torbat Heydarieh, Iran.*
[3] *Department of Computer Engineering, University of Mazandaran, Babolsar, Iran.*

Article Info

Article History:
Received 20 September 2025
Revised 15 November 2025
Accepted 31 December 2025

DOI:10.22044/jadm.2025.16791.2817

Keywords:
Speech Emotion Recognition; Visibility Graph; Feature Selection; Classification; Soft Voting.

**Corresponding author: z.mohammadpoory@shahroodut.ac.ir (Z. Mohammadpoury).*

Abstract

Nowadays, the recognition of emotions using speech signals has gained popularity because of its numerous applications in fields such as medicine, online marketing, search engines, education systems, criminal investigations, traffic collisions, and more. Many researchers have adopted different methodologies to improve emotion classification accuracy using speech signals. This study presents a novel time-series-to-graph transformation framework for speech emotion recognition. Speech signals were segmented into overlapping windows, each converted into a graph from which 16 structural features were extracted. Significant features were then selected via Minimum Redundancy Maximum Relevance (mRMR) and used to train four classifiers: random forest (RF), linear discriminant analysis (LDA), support vector machine (SVM), and k-nearest neighbors (KNN). Finally, a soft-voting ensemble strategy was employed to integrate their predictions, yielding improved classification performance. The proposed method achieved the highest sensitivity, specificity, and accuracy for the SAVEE database: 83.57%, 98.93%, and 98.16%, respectively. Similarly, for the EmoDB database, the highest values were 94.47%, 99.09%, and 98.40%, respectively. We also compared our results with other methods and found that our approach outperformed state-of-the-art techniques in emotion classification.

1. Introduction

Despite recent advancements in speech emotion recognition (SER) systems [1, 2], establishing reliable models capable of interpreting human emotions through speech remains an ongoing research challenge. SER systems play an essential role in various applications, including human–computer interaction, healthcare [3], emergency monitoring [4, 5], and engineering [6]. These systems help identify the cognitive, affective, and emotional states of speakers by analyzing expressive patterns embedded in speech.

Although speech is a natural communication medium that reflects human reactions, behaviors, and emotions, its processing is highly complex. Emotional content—such as happiness, anger, fear, sadness, and joy—introduces intricate fluctuations into speech signals. Furthermore, variations across speakers and languages complicate the design of robust feature representations [7]. Numerous algorithms based on time- and frequency-domain features have been developed to detect emotional states; however, further research is needed due to

the high dimensionality and nonlinear nature of speech signals, which make emotion monitoring both challenging and computationally demanding [8]. Therefore, developing accurate computational models can be highly beneficial for interpreting emotional responses. Motivated by this, the present study aims to design an SER system that focuses on extracting meaningful and discriminative features from speech signals.

Human speech is a nonlinear physiological signal in which glottal airflow dynamics reflect complex interactions of neural, muscular, and anatomical processes [9]. These dynamics are influenced by autonomic regulation and reactive behaviors to environmental stimuli [10]. As a result, nonlinear and dynamic approaches have gained increasing attention for analyzing speech characteristics [11, 12]. Among these, time-series-to-graph transformation techniques have emerged as powerful tools due to their capability to capture self-similarity and structural patterns within the signal [13]. The visibility graph (VG) is a prominent method that converts time series into graphs based on geometric visibility rules. The resulting adjacency matrix encodes local and global properties of the signal, effectively capturing its nonlinear complexity [14].

VG-based methods have been successfully applied in diverse fields such as biological signal processing [15, 16, 19], robotics [17], and architectural design [18]. In speech processing, the VG approach offers a strong advantage due to the inherently nonlinear and non-stationary nature of speech production mechanisms, influenced by airflow turbulence and neuromuscular interactions in the vocal tract [20]. Our earlier work demonstrated the potential of VG in differentiating speech patterns of Alzheimer's patients from healthy speakers [13]. Building on this foundation, the present study introduces a novel set of VG-based features for SER.

The primary objective of this study is to develop an effective speech emotion recognition system based on a time-series-to-graph transformation framework. The proposed approach leverages the visibility graph algorithm to model the nonlinear and non-stationary characteristics of speech signals and extract a comprehensive set of 16 graph-based structural features. These features are further refined using the mRMR method to identify the most informative attributes for emotion classification. The selected features are evaluated using four classifiers—RF, SVM, LDA, and KNN—and their complementary strengths are integrated through a soft-voting ensemble strategy to enhance robustness and classification performance. This framework aims to provide a discriminative and computationally efficient solution for identifying emotional states from speech signals.

The remainder of this paper is structured as follows: Section 2 reviews related work on SER and VG algorithms. Section 3 presents the proposed SER framework and its components. Sections 4 and 5 detail the experimental results and performance comparisons with other SER systems. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Related work

This section presents a brief background review of existing SER systems and VG applications, which motivated the proposed method of this study.

2.1. SER systems

In recent years, many studies have focused on extracting features from speech signals and leveraging them to achieve high SER accuracy. SER systems have been increasingly developed due to their important role in analyzing emotional states [21]. One of the main challenges in SER systems is eliciting informative data from speech signals. Therefore, selecting an appropriate feature extraction procedure is highly significant. Several studies have developed different sets of audio features by exploiting speech signal representations in the frequency and time domains [22]. In particular, promising emotion recognition results from actor-based datasets using spectrogram features were reported in [23, 24]. Nevertheless, no further examination has been conducted for free-context speech signals, which constitutes a major limitation.

Another widely utilized feature for SER is the Mel-frequency cepstral coefficients (MFCCs), which have proven effective in the automatic analysis and recognition of emotional states [25]. MFCCs are associated with the linear model of sound generation and the mapping of psychoacoustic frequency to the Mel scale [26]. Other popular features in this field include pitch, formant frequencies, energy, linear predictive spectrum coding, and Mel-energy spectrum dynamic coefficients (MESDC) [27]. Furthermore, some studies have employed various classification methods, including SVM, Hidden Markov Models, Multi-Layer Perceptron, and KNN, for emotion recognition [28–31]. For instance, Pan et al. combined MESDC, MFCC, and energy features of speech signals with SVM to recognize emotions [32]. Shegokar and Sircar [33] extracted prosodic

features based on continuous wavelet transform and used SVM for automatic SER.

Recent deep learning methods for SER utilize spectral representations and various feature sets, showing significant efficiency gains. However, they require expertise in training, large datasets, and deep learning knowledge. These techniques have gained attention for addressing these challenges, offering advantages over traditional methods by automatically identifying complex structures and features without manual extraction [21]. For example, Wang et al. [1] proposed an SER system called TF-Mix, employing three feature extraction techniques based on transformer architecture, Long Short-Term Memory (LSTM), and convolutional neural networks (CNN). The method was evaluated on different databases, such as SAVEE and RAVDESS, achieving accuracies of 99.857% and 87.513%, respectively. Poorna et al. [34] introduced a hybrid deep learning model using CNN and BiLSTM with multiple attention mechanisms, evaluated on the Amritaemo_Arabic, IEMOCAP, and Emo-DB datasets, achieving average accuracies of 95.80%, 67.85%, and 94.62%, respectively.

2.2. VG applications

Recently, some researchers have applied VG algorithms for various purposes and obtained favorable results. For example, Nasrolahzadeh et al. [35] adopted LSTM networks based on VG to distinguish heart rate variability during meditation. Ahmadlou et al. [36] applied VG analysis to electroencephalogram time series and its four sub-bands for the diagnosis of Alzheimer's disease. Mohammadpoory et al. [37] used weighted VG entropy to discriminate among healthy, ictal, and interictal groups in automatic epileptic seizure detection, achieving a diagnostic accuracy of 97%. Based on the VG method, they also proposed algorithms for mapping electrocorticogram time series to graphs capable of diagnosing epileptic seizures in rats [38]. Nasrolahzadeh et al. applied VG network analysis to heart rate time series to investigate the underlying mechanism of heart rate fluctuations during meditation from the viewpoint of the complex network hypothesis [39]. In another study [40], VG-based parameters were introduced to discriminate subjects with four different grades of diabetic retinopathy from fundus images.

3. Methodology

The proposed speech emotion recognition method is based on the VG approach and begins with preprocessing, where the input speech signal is segmented into fixed-size overlapping windows. Each window is then transformed into a visibility graph, from which a set of VG-based features is extracted. After computing these features, a feature selection technique is applied to identify the most informative ones. The selected features are subsequently used for classification, and finally, a soft-voting strategy is employed to generate the final decision. The proposed method was evaluated on two datasets. In the following sections, the datasets are first introduced, and then each stage of the proposed framework is described in detail.

3.1. Data Description

This paper employs two benchmark SER databases, which are described in detail below.

- dSAVEE: The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset is a recognized multimodal resource for emotion recognition, featuring 480 audio-visual recordings from four male English-speaking actors. They express seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality. This diversity in emotional expression makes SAVEE valuable for studying emotional cues in speech [41].
- EmoDB: The EmoDB database, developed by the Technical University of Berlin, contains recordings of 10 German actors expressing seven emotions—anger, boredom, disgust, fear, happiness, sadness, and neutral—through 10 semantically neutral sentences. This design isolates emotional context to vocal nuances, ensuring phonetic diversity across various German phonemes [42].

3.2. Preprocessing

At the first stage, the speech signals are preprocessed. To reduce variations between different speakers, the signals are normalized to zero mean and unit variance. A Voice Activity Detection (VAD) method is then applied to remove silence at the beginning and end of each signal. Finally, each signal is split into fixed-size overlapping windows.

3.3. Feature extraction using VG

The objective of the feature extraction phase is to identify crucial characteristics that can distinguish emotions from speech signals through VG-based features. VG was introduced by Lacasa for the transformation of time series into graphs [14]. In this method, each sample of the time series is represented by a node of the graph, known as the VG. The presence of an edge between two nodes implies that the corresponding time samples can "see" each other. The algorithm is based on

mapping time series X to its VG while preserving the time characteristic of the series. Assume that the jth data point of a time series is X_i. Then, two vertices X_m and X_n of the graph are connected by a bidirectional edge if the following inequality holds [13]:

$$x_{m+k} < x_n + \left(\frac{n-(m+k)}{n-m}\right)\left(x_m - x_n\right) \forall k \in Z^+; k < n-m \quad (1)$$

Figure 1 demonstrates the procedure of transforming time series into its VG. After transforming signal windows into VGs, some features were extracted from these graphs. Consider $A = \left[a_{ij}\right]_{N\times N}$ and $K = k(i)_{i=1,\ldots,N}$ which are assumed to be the adjacency matrix and degree sequence (DS) of graph network, respectively. In which N denotes the number of nodes in the VG. The existence of a connection between two nodes i and j make $b_{ij} = 1$. If there is no connection between two nodes i and j, it leads to $b_{ij} = 0$. In this work, sixteen different features were extracted from the graph as follows [40]:

The values of DS are expressed in terms of maximum, minimum, mean, mode, median, and standard deviation. Additionally, another feature is obtained by dividing the maximum value of DS by the median. Moreover, nine other features were extracted from the adjacency matrix of VG, namely: characteristic path length, global efficiency, local efficiency, average clustering coefficient, assortativity coefficient, graph entropy, graph index complexity, eccentricity, and radius of the graph [42].

In this study, all of the aforementioned features were estimated using MATLAB.

3.4. Feature Selection

Feature selection is essential in machine learning for improving accuracy, reducing overfitting, and enhancing interpretability [20]. Common approaches include filter, wrapper, and embedded methods, which evaluate feature importance and select optimal subsets during training. In this study, we employed the mRMR method to extract informative features. mRMR selects features that are highly relevant to the target variable while minimizing redundancy with other features. It ranks features based on mutual information with the target (significance) and with other features (redundancy), iteratively choosing those that maximize relevance and minimize overlap until the desired subset is obtained. Despite its computational cost, mRMR effectively balances significance and redundancy, making it valuable for high-dimensional data [43].

3.5. Classification method

Classification is one of the most significant components of diagnostic systems, as its performance directly affects the system's accuracy. In this paper, four well-known classifiers—RF, KNN, SVM, and LDA—are employed as follows.

RF [15] is an ensemble learning method that uses multiple decision trees and bagging to improve accuracy and reduce overfitting. It is robust to outliers and can handle imbalanced data effectively.

KNN [12] is a simple yet effective method that classifies data based on the majority vote of the k nearest neighbors in the feature space.

SVM [12] is a flexible classifier that finds the optimal hyperplane for separating classes and uses kernel functions to handle non-linear data.

LDA [44] is a statistical model that projects data onto a lower-dimensional space to maximize class separation, assuming that the classes are normally distributed with equal covariances.

As mentioned earlier, we applied the soft voting method to improve the performance of the proposed framework. Soft voting [45] is an ensemble learning technique in which classifiers output probability distributions over all classes, and the final decision is made by averaging these probabilities. This approach integrates the strengths of multiple classifiers and considers the confidence level of their predictions, making it more robust than hard voting, where all votes are weighted equally. In our case, the four classifiers each provided probability estimates for every data point, and the averaged result determined the final class label. Such a strategy enhances accuracy, stability, and overall recognition performance, assuming that the classifiers are properly calibrated and generate reliable probability scores.

Two approaches were employed to classify emotions using VG-based features. The first approach involved concatenating all features from every frame of a signal to create a feature vector, which was then used for training and testing by a classifier. The second approach entailed extracting statistical metrics—including mean, variance, skewness, kurtosis, minimum, and maximum—from each feature across all frames of a signal for classification purposes.

4. Experimental set up

To assess the proposed approach, all experiments were performed using MATLAB R2024a on a

system powered by a 2.3 GHz Intel Core i7 processor and 24 GB RAM.

In this paper, we used 5-fold cross-validation, and the evaluation procedure was repeated five times to obtain more reliable and stable results. Repeating the cross-validation reduces the effect of random data splitting and provides a more robust estimate of the model's generalization performance. Additionally, the data were split based on subjects rather than frames, meaning that samples from the same individual were not shared between the training and test sets. This subject-independent evaluation prevents data leakage and leads to a more realistic and fair assessment of the model's performance, especially in practical scenarios where the model is applied to unseen subjects.

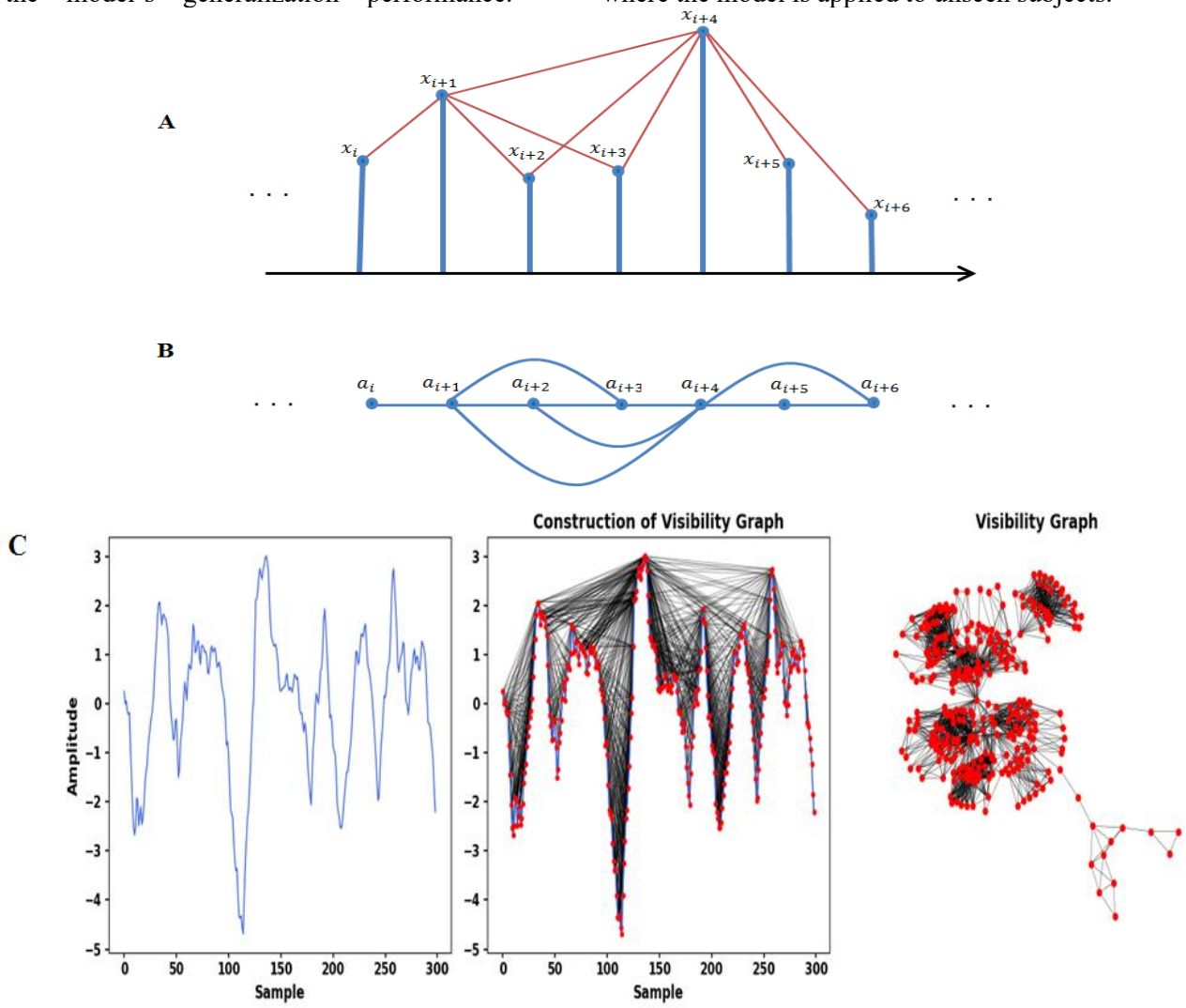

Figure 1. (A) An example of a speech time series with seven data points. (B) The VG constructed from the time series, where every two data points that can "see" each other are connected by a red line. (C) An example of a 300-point speech time series from the EmoDB dataset and its corresponding VG construction.

As mentioned above, prior to feature extraction, each speech signal was segmented into N sample frames with M sample overlaps. Various combinations of frame and overlap lengths (N-M) were employed, including 500-150, 750-250, 1000-300, 1400-400, and 2000-1000 samples. Optimal performance was observed with 1400-400 for the SAVEE dataset and 750-250 for the EmoDB database. Therefore, the results for frame lengths and overlaps other than those specified were omitted from this paper. Since the most time-consuming part of the proposed method is the construction of the VG and the extraction of graph-based features, we report its runtime. The average runtime per speech segment for the optimal frame lengths was approximately 71.22 seconds for 1400sample frames in the SAVEE dataset and 25.91 seconds for 750 sample frames in the EmoDB dataset, demonstrating that the proposed method is computationally efficient for practical use.

To standardize signal lengths and feature numbers, only the initial 60 frames of each signal were chosen for further analysis. For signals with fewer

than 60 frames, the final frames were duplicated to reach 60 frames. Subsequently, 16 VG features were extracted from all frames and classified using four classifiers through two approaches. The combined outcomes of the classifiers were aggregated using the soft voting technique to enhance the efficacy of the proposed approach. As previously stated, the SAVEE dataset comprises four files featuring speech signals from different speakers. Each file contains 120 speech signals across seven emotions, with 15 segments per non-neutral emotion and 30 for neutral (15 selected).

The EmoDB database comprises 535 signals, with 127, 62, 79, 81, 71, 46, and 69 segments corresponding to anger, sadness, neutral, boredom, happiness, disgust, and anxiety/fear emotions, respectively. To address class imbalance and enhance classifier performance, the training set was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) [46]. Subsequently, Range Normalization [12] was applied to scale data values within the range [0, 1] for the features.

Three measures were used for evaluation of the proposed method: Accuracy, Sensitivity, and specificity [47].

In the first approach, the classifiers were provided with 60×16 = 960 input features. The mRMR feature selection was employed to reduce the feature space and identify the most discriminative features. The mRMR scores for the SAVEE dataset and the 1th fold are illustrated in Figure 2. A sharp decline in the mRMR score occurs after the 160th feature; therefore, only the first 200 features are visualized. For each classifier, different numbers of top-ranked features (ranging from 1 to 160) were used as inputs, and the best results were reported. The findings indicate that 32 selected features with the highest mRMR scores achieved the best performance in this fold. In each fold, a different number of features and a different set of features produced the best results.

Figure 3 presents the mRMR scores for the EmoDB dataset and the 1th fold. Similar to the previous case, a noticeable drop in the mRMR score appears after the 190th feature, which is why only the first 200 features are displayed. Classifiers were then constructed using varying numbers of top-ranked features (ranging from 1 to 190). The results show that using 25 selected features with the highest mRMR scores led to the best results in this fold. Across the folds, both the number of selected features and the specific feature set varied in producing the best performance.

Tables 1 and 2 present the average confusion matrixes obtained from the first approach applied to the SAVEE and EmoDB databases.

In the second approach, the classifiers were initially provided with 16×6=96 input features, and the MRMR method was employed to reduce the feature set to a more manageable size while retaining the most relevant features. A procedure similar to the first approach was applied for selecting the optimal features and performing classification for each fold and classifier. Tables 3 and 4 present the average confusion matrices obtained over 25 folds by the two proposed approaches for the two mentioned datasets.

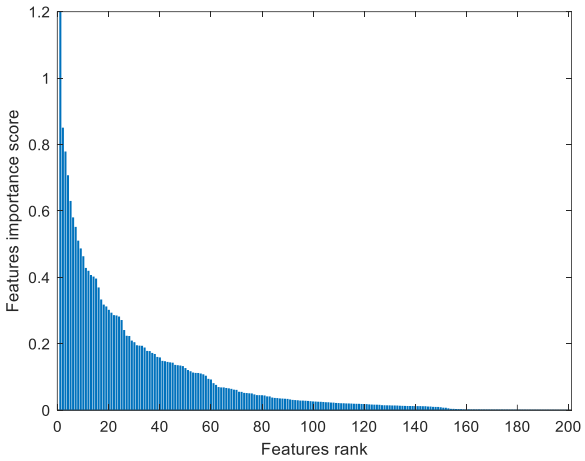

Figure 2. The mRMR scores of 200 features for the SAVEE database and the first approach.

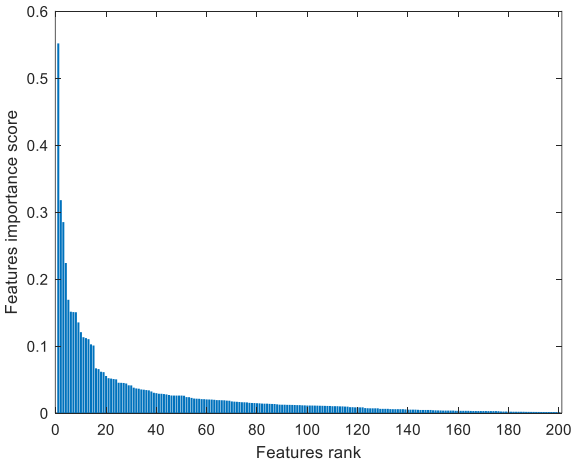

Figure 3. The mRMR scores of 200 features for the EmoDB database and the first approach.

Table 5 summarize the performance of the proposed method in terms of accuracy, sensitivity, and specificity across both approaches and datasets. The reported sensitivities, specificities, and accuracies are macro-averaged across all emotion classes. Performance metrics are reported as mean ± standard deviation (SD) over all 25 folds. Based on the results in Table 5, the second approach performs better than the first approach on both the SAVEE and EmoDB databases. For the

SAVEE dataset, sensitivity increases from 87.62 ± 3.1% to 93.57 ± 2.26%, indicating that the second approach detects target samples more accurately and with less variation. At the same time, specificity improves from 97.94 ± 1.24% to 98.93 ± 1.01%, and accuracy rises from 96.24 ± 1.97% to 98.16 ± 1.19%. A similar pattern is observed in the EmoDB dataset, where sensitivity increases from 90.35 ± 3.41% to 94.47 ± 2.54%, while specificity improves from 98.39 ± 1.45% to 99.09 ± 0.21%, and accuracy rises from 97.22 ± 1.76% to 98.40 ± 0.65%. Overall, the higher mean values and the generally lower standard deviations indicate that the second approach is not only more accurate but also more stable and reliable across different runs.
The p-values reported in the table were obtained using t-tests to assess whether the differences in performance between the first and second approaches are statistically significant. These tests compare the results across multiple folds of cross-validation to determine if the observed improvements are likely due to the proposed method rather than random variation. In our results, all p-values are less than 0.05, indicating that the increases in sensitivity, specificity, and accuracy achieved by the second approach are statistically significant. This confirms that the improvements are real, meaningful, and not caused by chance. We also performed paired t-tests to evaluate whether the performance improvements achieved by the proposed ensemble method are statistically significant when compared with individual classifiers (RF, SVM, LDA, and KNN), as well as the second approach without feature selection. The corresponding p-values for all pairwise comparisons are reported in Tables 6 and 7 for both datasets. The results indicate that combining classifiers through soft voting and applying feature selection consistently leads to better performance than using individual classifiers alone and the second approach without MRMR. In most cases, these improvements are statistically significant ($p < 0.05$).

Table 1. Confusion matrix for the first approach using selected features by the MRMR method, on the SAVEE database.

classes	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral	Total
Anger	51	2	2	1	1	3	0	60
Disgust	2	54	2	0	1	1	0	60
Fear	2	2	51	1	1	3	0	60
Happiness	1	0	1	53	2	2	1	60
Sadness	2	1	1	1	51	2	2	60
Surprise	1	1	1	2	1	54	0	60
Neutral	1	1	0	2	1	1	54	60
Total	60	61	58	60	58	66	57	420

Table 2. Confusion matrix for the first approach using selected features by the MRMR method, on the EmoDB database.

classes	Anger	Disgust	Anxiety/Fear	Happiness	Sadness	Boredom	Neutral	Total
Anger	115	3	4	2	1	2	0	127
Disgust	1	63	1	1	0	3	0	69
Anxiety/Fear	1	1	42	2	0	0	0	46
Happiness	1	0	1	64	3	2	0	71
Sadness	1	2	0	1	56	2	0	62
Boredom	2	1	4	2	1	71	0	81
Neutral	0	1	0	2	3	1	72	79
Total	121	71	52	74	64	81	72	535

Table 3. Confusion matrix for the second approach using selected features by the MRMR method, on the SAVEE database.

Classes	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral	Total
Anger	56	1	0	2	1	0	0	60
Disgust	1	55	1	0	1	2	0	60
Fear	0	1	55	1	2	1	0	60
Happiness	0	0	0	57	0	1	2	60
Sadness	1	1	0	0	56	2	0	60
Surprise	0	0	1	0	1	57	1	60
Neutral	0	0	0	1	0	2	57	60
Total	58	58	57	61	61	65	60	420

Table 4. Confusion matrix for the second approach using selected features by the MRMR method, on the EmoDB database.

classes	Anger	Disgust	Anxiety/Fear	Happiness	Sadness	Boredom	Neutral	Total
Anger	120	1	4	1	0	1	0	127
Disgust	0	65	0	3	0	1	0	69
Anxiety/Fear	0	0	43	1	2	0	0	46
Happiness	0	0	0	70	0	1	0	71
Sadness	0	1	0	0	60	1	0	62
Boredom	1	2	3	1	0	74	0	81
Neutral	0	0	2	1	3	0	73	79
Total	121	69	52	77	65	78	73	535

Table 5. The evaluation measures (mean±SD) of the proposed method for SAVEE and EmoDB databases and P-values for both approaches.

Approach / Database	The first approach/ SAVEE	The second approach / SAVEE	P-value	The first approach / EmoDB	The second approach / EmoDB	P-value
Sensitivity%	87.62±3.1	93.57±2.26	0.01	90.35±3.41	94.47±2.54	0.02
Specificity%	97.94±1.24	98.93±1.01	0.03	98.39±1.45	99.09±0.21	0.04
Accuracy%	96.24±1.97	98.16±1.19	0.01	97.22±1.76	98.40±0.65	0.01

Table 6. Performance comparison of the proposed ensemble method and individual classifiers on the SAVEE database with p-values.

Method	Accuracy	P-value with the second approach
RF+MRMR	91.32±4.56	0.04
SVM+MRMR	89.64±4.78	0.003
LDA+MRMR	88.48±2.13	0.05
KNN+MRMR	88.39±2.98	0.02
Ensemble without MRMR	92.53±3.11	0.03

Table 7. Performance comparison of the proposed ensemble method and individual classifiers on the EmoDB database with p-values.

Method	Accuracy	P-value with the second approach
RF+MRMR	90.87±4.15	0.05
SVM+MRMR	90.52±5.46	0.01
LDA+MRMR	89.65±2.14	0.001
KNN+MRMR	89.86±1.98	0.002
Ensemble without MRMR	91.54±1.96	0.04

5. Discussion and future work

In this study, we proposed a speech emotion recognition system based on VG features. The VG-based approach offers a unique perspective on the temporal dynamics of speech signals, capturing intricate patterns that traditional time- or frequency-domain methods might overlook [13]. The extracted VG features were further refined using the mRMR method, ensuring the selection of the most informative features with minimal redundancy. We employed four classifiers—RF, KNN, SVM, and LDA—to evaluate the efficacy of the selected features. The classifiers' outputs were then combined using a soft voting method, enhancing the robustness and accuracy of the final decision. Two feature classification approaches were evaluated on the SAVEE and EmoDB databases. For SAVEE, a frame length of 1400 samples with an overlap of 400 samples yielded the best results, while for EmoDB, a frame length of 750 samples with an overlap of 250 samples was optimal. The use of range normalization and SMOTE for balancing the training data further improved classifier performance.

The findings from this study underscore the effectiveness of the VG-based feature extraction method combined with mRMR for SER. The superior performance of the second approach, particularly in terms of accuracy and feature efficiency, suggests that leveraging statistical measures for feature classification is highly beneficial. This can be attributed to several factors. Concatenating all frame-level features generates high-dimensional vectors, which increase the risk of overfitting given the limited size of SER datasets. In contrast, statistical measures such as mean, variance, skewness, kurtosis, minimum, and maximum substantially reduce dimensionality and produce more compact and generalizable representations. Since emotion labels are typically assigned at the utterance level rather than the frame level, aggregated statistics align more naturally with the ground-truth labels by filtering out short-term variability and noise. Statistical descriptors also offer robustness to temporal misalignments and redundancies inherent in frame-level features, allowing the classifier and feature selection algorithm to operate more effectively. Overall, distribution-based summarization of VG features captures the essential characteristics of emotional expressions in a more reliable and discriminative manner.

Our findings demonstrate the proposed system's robustness and generalizability across diverse

databases. Consistent performance improvements in the SAVEE and EmoDB datasets show its reliability in distinguishing emotional states in speech. The integration of multiple classifiers and soft voting further enhances accuracy and resilience.

In terms of classification accuracy, our results indicate a progression relative to other techniques [1, 2, 48–50], even though we only used 20 and 21 features. For comparison purposes, we only present the results of the current study alongside our previous studies that used the same database. Table 8 briefly compares the current study with previous studies regarding the effectiveness of diagnosis. The proposed approach outperforms the others. Therefore, we can confidently state that the present study demonstrates more efficient performance than previous works. Since the VG method for speech time series is relatively new and has shown efficiency in speech processing, this approach could be useful for other speech processing applications, such as automatic speaker recognition.

Future research could explore the application of this VG-based method to real-time emotion recognition systems, which would have significant implications for interactive voice response systems and clinical settings for monitoring emotional well-being. Additionally, extending the method to multimodal emotion recognition systems that incorporate facial expressions, gestures, and physiological signals could further enhance its robustness and applicability.

Table 8. Comparison of the present study with previous studies.

Reference	Year	Features	Classifier	Database	Accuracy (%)
Mishra et al. [48]	2023	TF-based permutation entropy (TFPE) feature	SVM, RF, DT, and KNN	SAVEE/EmoDB	68.78/77.2
Xie et al. [49]	2023	Constant-Q spectrogram-based histogram of oriented gradients, openSMILE, and wavelet packet decomposition-based features	Random forest and Grey wolf optimization	SAVEE/EmoDB	88.79/95.29
Wang et al. [1]	2024	CNNs, LSTMs, and Transformer architecture-based features	CNN, BiLSTM-FCN, and BiLSTM-Transformer-FCN	SAVEE	86.233
Mishra et al. [50]	2024	Spectrogram, MFCC, mel-spectrogram	DNN+ CNN	SAVEE	80.99
Mishra et al. [2]	2025	MRHT-based entropy feature	DNN	SAVEE/EmoDB	83.48/89.67
Proposed method	2025	VG based features	Soft voting combining KNN, LDA, SVM, RF	SAVEE/EmoDB	98.16/98.40

6. Conclusion

In this paper, we propose a VG-based system for SER using the SAVEE and EmoDB databases. Results demonstrate that the second approach, which employs statistical measures for feature classification, consistently outperforms the first approach by achieving higher accuracies with fewer features, reaching 98.16% on SAVEE with 21 features and 98.40% on EmoDB with 20 features. In comparison, the first approach obtained 96.24% on SAVEE with 27 features and 97.22% on EmoDB with 30 features. Sensitivity and specificity were also higher for the second approach, achieving 93.57% and 98.93% for SAVEE, and 94.47% and 99.09% for EmoDB. These results indicate that the second approach is not only more accurate but also more efficient and computationally feasible. Overall, the proposed VG-based framework offers a robust and innovative solution for emotion recognition, with applications in human-computer interaction and mental health monitoring. For future work, we aim to apply this method to larger and more diverse emotional speech datasets, explore real-time implementations, and integrate graph-based features with deep learning models to further enhance performance and robustness.

References

[1] M. Wang, H. Ma, Y. Wang, and X. Sun, "Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion," *Appl. Acoust.*, vol. 218, p. 109886, 2024.

[2] S. P. Mishra, P. Warule, and S. Deb, "Speech emotion recognition using multi resolution Hilbert transform based spectral and entropy features," *Appl. Acoust*., vol. 229, p. 110403, 2025.

[3] H. Wang, Y. Liu, X. Zhen, and X. Tu, "Depression speech recognition with a three-dimensional convolutional network*," Front. Hum. Neurosci*., vol. 15, p. 713823, 2021.

[4] M. Bojanić, V. Delić, and A. Karpov, "Call redistribution for a call center based on speech emotion recognition," *Appl. Sci*., vol. 10, no. 13, p. 4653, 2020.

[5] T. Deschamps-Berger, L. Lamel, and L. Devillers, "End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings," *in 2021 9th Int. Conf. Affective Comput. Intell. Interact. (ACII)*, pp. 1–8, IEEE, 2021.

[6] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," *in ICASSP 2021*, pp. 5734–5738, IEEE, 2021.

[7] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit*ion., vol. 44, no. 3, pp. 572–587, 2011.

[8] J. Kacur, B. Puterka, J. Pavlovicova, and M. Oravec, "On the speech properties and feature extraction methods in speech emotion recognition," *Sensors,* vol. 21, no. 5, p. 1888, 2021.

[9] Z. Mohammadpoory, M. Nasrolahzadeh, S. A. Amiri, and J. Haddadnia, "A Non-invasive Approach for Early Alzheimer's Detection Through Spontaneous Speech Analysis Using Deep Visibility Graphs," *Cogn. Comput*., vol. 17, no. 1, pp. 42, 2025.

[10] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[11] M. Nasrolahzadeh, Z. Mohammadpoory, and J. Haddadnia, "Weighted Visibility Graph-based Deep Complex Network Features: New Diagnostic Spontaneous Speech Markers of Alzheimer's Disease," *Physica D*, vol. 476, p. 134681, 2025.

[12] F. Mohammady, S. Asadi Amiri, and Z. Mohammadpoory, "Leveraging segmentation and visibility graph analysis to enhance motor imagery classification in EEG signals," *Cogn. Comput*., vol. 18, no. 1, p. 8, Dec. 2026.

[13] M. Nasrolahzadeh, Z. Mohammadpoory, and J. Haddadnia, "Indices from visibility graph complexity of spontaneous speech signal: An efficient nonlinear tool for Alzheimer's disease diagnosis," *Chaos Solitons Fractals*, vol. 174, p. 113829, 2023.

[14] L. Lacasa, B. Luque, J. Luque, and J. C. Nuno, "The visibility graph: A new method for estimating the Hurst exponent of fractional Brownian motion," *Europhys. Lett.*, vol. 86, no. 3, p. 30001, 2009.

[15] Z. Mohammadpoory, M. Nasrolahzadeh, S. A. Amiri, "Classification of healthy and epileptic seizure EEG signals based on different visibility graph algorithms and EEG time series," *Multimed. Tools Appl*., vol. 83, pp. 2703–2724, 2024.

[16] M. Nasrolahzadeh, Z. Mohammadpoory, and J. Haddadnia, "The visibility graph analysis of heart rate variability during chi meditation and Kundalini Yoga techniques," *Healthcare Analytics*, vol. 4, p. 100253, 2023.

[17] Y. You, C. Cai, and Y. Wu, "3D visibility graph based motion planning and control," in Proc. *5th Int. Conf. Robotics Artif. Intell*., pp. 48–53, 2019.

[18] T. Varoudis and S. Psarra, "Beyond two dimensions: architecture through three dimensional visibility graph analysis," *J. Space Syntax*, vol. 5, no. 1, pp. 91–108, 2014.

[19] Z. Mohammadpoory, M. Nasrolahzadeh, S. A. Amiri, "Patient-independent epileptic seizure detection using weighted visibility graph features and wavelet decomposition," *Multimed. Tools Appl*., pp. 1–25, 2025.

[20] M. Nasrolahzadeh, Z. Mohammadpoory, and J. Haddadnia, "A novel method for early diagnosis of Alzheimer's disease based on higher-order spectral estimation of spontaneous speech signals," *Cogn. Neurodyn*., vol. 10, no. 6, pp. 495–503, 2016.

[21] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021.

[22] B. Schuller and A. Batliner, Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. *John Wiley & Sons*, 2013.

[23] M. Papakostas, G. Siantikos, T. Giannakopoulos, E. Spyrou, and D. Sgouropoulos, "Recognizing emotional states using speech information," *in GeNeDis 2016: Geriatrics*, pp. 155–164, Springer, 2017.

[24] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[25] K. Tomba, J. Dumoulin, E. Mugellini, O. Abou Khaled, and S. Hawila, "Stress detection through speech analysis," *in ICETE 2018*, pp. 560–564, 2018.

[26] L. Vignolo, H. Rufiner, and D. Milone, "Multi-objective optimisation of wavelet features for phoneme recognition," *IET Signal Process*., Available: http://digital-library.theiet.org/content/journals/10.1049/iet-spr.

[27] K. Aghajani and I. E. Paeen Afrakoti, "Speech emotion recognition using scalogram based deep structure," *Int. J. Eng*., vol. 33, no. 2, pp. 285–292, 2020.

[28] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit*., vol. 44, no. 3, pp. 572–587, 2011.

[29] S. Taran, "A nonlinear feature extraction approach for speech emotion recognition using VMD and TKEO," *Appl. Acoust*., vol. 214, p. 109667, 2023.

[30] R. K. Srivastava and D. Pandey, "Speech recognition using HMM and Soft Computing," *Mater. Today: Proc*., vol. 51, pp. 1878–1883, 2022.

[31] J. M. López-Gil and N. Garay-Vitoria, "Assessing the effectiveness of ensembles in Speech Emotion Recognition: Performance analysis under challenging scenarios," *Expert Syst. Appl*., vol. 243, p. 122905, 2024.

[32] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *Int. J. Smart Home*, vol. 6, no. 2, pp. 101–108, 2012.

[33] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," *in 2016 10th Int. Conf. Signal Process. Commun. Syst. (ICSPCS),* pp. 1–8, IEEE, 2016.

[34] S. S. Poorna, V. Menon, and S. Gopalan, "Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition," *Biomed. Signal Process. Control*, vol. 100, p. 106967, 2025.

[35] M. Nasrolahzadeh, Z. Mohammadpoory, "A novel method for distinction heart rate variability during meditation using LSTM recurrent neural networks based on visibility graph," *Biomed. Signal Process. Control,* vol. 90, p. 105822, 2024.

[36] M. Ahmadlou, H. Adeli, and A. Adeli, "New diagnostic EEG markers of the Alzheimer's disease using visibility graph," *J. Neural Transm*., vol. 117, pp. 1099–1109, 2010.

[37] Z. Mohammadpoory, M. Nasrolahzadeh, and J. Haddadnia, "Epileptic seizure detection in EEGs signals based on the weighted visibility graph entropy," *Seizure*, vol. 50, pp. 202–208, 2017.

[38] Z. Mohammadpoory, M. Nasrolahzadeh, N. Mahmoodian, M. Sayyah, and J. Haddadnia, "Complex network based models of ECoG signals for detection of induced epileptic seizures in rats," *Cogn. Neurodyn*., vol. 13, pp. 325–339, 2019.

[39] M. Nasrolahzadeh, Z. Mohammadpoory, and J. Haddadnia, "Analysis of heart rate signals during meditation using visibility graph complexity," *Cogn. Neurodyn*., vol. 13, pp. 45–52, 2019.

[40] Z. Mohammadpoory, M. Nasrolahzadeh, N. Mahmoodian, and J. Haddadnia, "Automatic identification of diabetic retinopathy stages by using fundus images and visibility graph method," *Measurement*, vol. 140, pp. 133–141, 2019.

[41] S. Haq and P. J. B. Jackson, "Speaker-dependent audio-visual emotion recognition," *in Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP)*, pp. 53–58, 2009.

[42] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Proc. Interspeech, Lisbon, Portugal*, pp. 1517–1520, 2005.

[43] S. Asadi Amiri, M. Nasrolahzadeh, Z. Mohammadpoory, A. Movahedinia, and A. Zare, "A Novel Method for Fish Spoilage Detection based on Fish Eye Images using Deep Convolutional Inception-ResNet-v2," *J. AI Data Min*., vol. 12, no. 1, pp. 105–113, 2024.

[44] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[45] L. Breiman, "Bagging predictors," *Mach. Learn*., vol. 24, no. 2, pp. 123–140, 1996.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res*., vol. 16, pp. 321–357, 2002.

[47] S. Moghani, H. Marvi, and Z. Mohammadpoory, "Valvular Heart Disease Classification through Hierarchical Decomposition via Matrix Factorization of Scalogram-Based Phonocardiogram Representations," *Journal of AI and Data Mining*, vol. 13, no. 3, pp. 369–378, Jul. 2025.

[48] S. P. Mishra, P. Warule, and S. Deb, "Chirplet transform based time frequency analysis of speech signal for automated speech emotion recognition," *Speech Commun*., vol. 155, p. 102986, 2023.

[49] J. Xie, M. Zhu, and K. Hu, "Fusion-based speech emotion classification using two-stage feature selection," *Speech Commun*., vol. 152, p. 102955, 2023.

[50] S. P. Mishra, P. Warule, and S. Deb, "Speech emotion classification using feature-level and classifier-level fusion," *Evolv. Syst*., vol. 15, no. 2, pp. 541–554, 2024.

یادگیری تجمعی برای تشخیص احساس از روی گفتار با استفاده از پویایی سیگنال مبتنی بر گراف

زینب محمدپوری[1,*]، مهدا نصرالله زاده[2] و سکینه اسدی امیری[3]

[1] دانشکده مهندسی برق، دانشگاه صنعتی شاهرود، شاهرود، ایران.

[2] گروه مهندسی برق، دانشگاه تربت حیدریه، تربت حیدریه، ایران.

[3] گروه مهندسی کامپیوتر، دانشگاه مازندران، بابلسر، ایران.

ارسال ۲۰۲۵/۰۸/۲۰؛ بازنگری ۲۰۲۵/۱۱/۱۵؛ پذیرش ۲۰۲۵/۱۲/۳۱

چکیده:

امروزه تشخیص احساس با استفاده از سیگنال‌های گفتاری به دلیل کاربردهای فراوان آن در حوزه‌هایی مانند پزشکی، بازاریابی آنلاین، موتورهای جستجو، سامانه‌های آموزشی، تحقیقات جنایی، تصادفات رانندگی و موارد دیگر، محبوبیت زیادی پیدا کرده است. پژوهشگران بسیاری روش‌های گوناگونی را برای بهبود دقت طبقه‌بندی هیجان با استفاده از سیگنال گفتار به‌کار گرفته‌اند.در این پژوهش، یک چارچوب نوین برای تبدیل سری زمانی به گراف به‌منظور تشخیص هیجان گفتار ارائه می‌شود. سیگنال‌های گفتاری ابتدا به پنجره‌های هم‌پوشان تقسیم شدند و سپس هر پنجره به یک گراف تبدیل گردید که از آن ۱۶ ویژگی ساختاری استخراج شد. در ادامه، ویژگی‌های معنادار با استفاده از روش حداقل افزونگی و حداکثر ارتباط انتخاب شدند و برای آموزش چهار طبقه‌بند شامل جنگل تصادفی، تحلیل تفکیکی خطی، ماشین بردار پشتیبان و k نزدیک‌ترین همسایه به‌کار رفتند. در نهایت، یک راهبرد تجمعی رأی‌گیری نرم برای ادغام پیش‌بینی‌های این طبقه‌بندها استفاده شد که موجب بهبود عملکرد طبقه‌بندی گردید. روش پیشنهادی در پایگاه داده SAVEE بالاترین مقادیر حساسیت، ویژگی و دقت را به‌ترتیب برابر با ۸۳.۵۷٪، ۹۸.۹۳٪ و ۹۸.۱۶٪ به‌دست آورد. همچنین در پایگاه داده EmoDB بیشترین مقادیر حساسیت، ویژگی و دقت به‌ترتیب برابر با ۹۴.۴۷٪، ۹۹.۰۹٪ و ۹۸.۴۰٪ گزارش شد. علاوه بر این، نتایج با سایر روش‌ها مقایسه گردید و مشخص شد رویکرد پیشنهادی عملکرد بهتری نسبت به روش‌های پیشرفته موجود در طبقه‌بندی احساس ارائه می‌دهد.

کلمات کلیدی: تشخیص احساس از روی گفتار، گراف پدیداری، انتخاب ویژگی، طبقه بندی، رای گیری نرم.