MESTRO JURNAL ILMIAH
Jurnal homepage: http://jurnalmestro.untag.ac.id

Acoustic Pattern Classification in Female Voice Using K-Nearest Neighbor with MFCC Feature Extraction

*Aris Rakhmadi[1], Joko Handoyo[2], Irma Yuliana[3], Dimara Hakim Kusuma[4]

[1)]Department of Informatics Engineering, Faculty of Communication Science and Informatics, Universitas Muhammadiyah Surakarta, Jawa Tengah, INDONESIA.
[2)]Department of Informatics, Ronggolawe College of Technology, Blora, Jawa Tengah, INDONESIA.
[3)]Department of Informatics Education, Faculty of Teaching and Education Training, Universitas Muhammadiyah Surakarta, Jawa Tengah, INDONESIA.
[4)]Department of Information Systems, Faculty of Technology and Sciences, Universitas Muhammadiyah Purwokerto, Jawa Tengah, INDONESIA.

INFORMASI ARTIKEL

NASKAH DITERIMA : 12 April 2026
DIREVISI : 25 Mei 2026
DISETUJUI : 25 Juni 2026

***KORESPODENSI PENULIS :**
aris.rakhmadi@ums.ac.id

DOI: 10.47685/mestro.v8i01.794

Hal. 1 - 11

Abstract

This study investigates acoustic pattern classification in female voice data using Mel-Frequency Cepstral Coefficient (MFCC) features and the K-Nearest Neighbor (KNN) algorithm. The analysis used the publicly available "MFCCs for Speech Emotion Recognition" dataset from Kaggle. A female-only subset containing 49,225 observations across eight predefined categories—angry, calm, disgust, fear, happy, neutral, sad, and surprise—was selected. Each observation was represented by 58 precomputed MFCC-based features, and the data were divided into 80% training and 20% testing subsets. The novelty of this study lies in its single-gender, female-only evaluation, which provides a more focused analysis of acoustic category separability without combining male and female voice characteristics. KNN was evaluated using accuracy, weighted precision, recall, F1-score, and a confusion matrix, while Support Vector Machine (SVM) was included as a comparative baseline. KNN achieved an accuracy of 87.75%, a weighted precision of 0.8792, a recall of 0.8775, and an F1-score of 0.8775, outperforming SVM, which obtained an accuracy of 80.85% and an F1-score of 0.8082. Category-level analysis showed that surprise and calm were recognized effectively, whereas happy, disgust, fear, neutral, and sad exhibited greater acoustic overlap. These findings demonstrate that KNN provides a competitive, lightweight baseline for classifying acoustic patterns in female voices, although further validation with cross-validation, temporal features, and advanced models is required.

Keywords: *Emotion Pattern Recognition, MFCC Feature Extraction, K-Nearest Neighbor, Female Acoustic Voice, Machine Learning*

Abstrak

Penelitian ini mengkaji klasifikasi pola akustik pada sinyal suara perempuan menggunakan fitur *Mel-Frequency Cepstral Coefficient* (MFCC) dan algoritma *K-Nearest Neighbor* (KNN). Analisis menggunakan dataset publik "MFCCs for Speech Emotion Recognition" yang diperoleh dari Kaggle. Subset khusus suara perempuan sebanyak 49.225 observasi dipilih dan dikelompokkan ke dalam delapan kategori, yaitu *angry*, *calm*, *disgust*, *fear*, *happy*, *neutral*, *sad*, dan *surprise*. Setiap observasi direpresentasikan menggunakan 58 fitur berbasis MFCC yang telah tersedia dalam dataset, kemudian data dibagi menjadi 80% data pelatihan dan 20% data pengujian. Kebaruan penelitian ini terletak pada evaluasi single-gender yang secara khusus menggunakan data suara perempuan sehingga analisis keterpisahan pola akustik tidak dipengaruhi oleh penggabungan karakteristik suara laki-laki dan perempuan. Performa KNN dievaluasi menggunakan akurasi, *weighted precision*, *recall*, *F1-score*, dan matriks konfusi, sedangkan *Support Vector Machine* (SVM) digunakan sebagai metode pembanding. KNN memperoleh akurasi 87,75%, weighted precision 0,8792, recall 0,8775, dan F1-score 0,8775, lebih tinggi dibandingkan SVM yang memperoleh akurasi 80,85% dan F1-score 0,8082. Analisis per kategori menunjukkan bahwa surprise dan calm dapat dikenali dengan baik, sedangkan happy, disgust, fear, neutral, dan sad menunjukkan tumpang tindih pola akustik yang lebih besar. Hasil ini menunjukkan bahwa KNN dapat menjadi metode dasar yang ringan dan kompetitif untuk klasifikasi pola akustik suara perempuan, meskipun validasi lebih lanjut menggunakan cross-validation, fitur temporal, dan model yang lebih kompleks masih diperlukan.

Kata kunci: Pengenalan Pola Emosi, Ekstraksi Ciri MFCC, K-Nearest Neighbor, Suara Akustik Wanita, Machine Learning

I. INTRODUCTION

Human voice signals carry rich information beyond spoken words, including variations in tone, intensity, and spectral characteristics that reflect underlying vocal patterns [1]. These acoustic properties can be analyzed computationally to understand differences in speech production and identify distinct patterns across categories [2], [3]. As a result, the analysis of acoustic

features in speech signals has become an important topic in the development of intelligent systems, particularly in applications related to voice-based interaction, behavioral analysis, and adaptive computing systems.

However, analyzing acoustic variations in speech signals is not a straightforward task. Multiple factors, such as individual vocal traits, recording environments, and speaking styles, influence speech characteristics. These factors introduce variability in the extracted features, which may reduce the effectiveness of classification models [4], [5]. In particular, overlapping acoustic patterns across different categories can make it difficult for machine learning algorithms to achieve consistent classification performance. Therefore, the selection of appropriate feature representations and classification methods is critical to acoustic pattern analysis [6], [7].

Among various feature extraction techniques, Mel-Frequency Cepstral Coefficients (MFCC) are widely used to represent the spectral properties of speech signals. MFCC features are designed to model the human auditory system, enabling them to capture important frequency components that characterize voice signals [8], [9], [10]. Due to their effectiveness in representing spectral envelopes, MFCC features have been extensively applied in various speech-related tasks, including voice analysis and pattern classification. Their ability to preserve essential acoustic information makes them suitable for distinguishing variations in voice characteristics.

In addition to feature representation, the choice of an appropriate classification algorithm plays a key role in determining the overall performance of an acoustic pattern recognition system. Previous speech emotion recognition studies have employed various machine learning and deep learning methods, including K-Nearest Neighbor, Support Vector Machine, Random Forest, convolutional neural networks, recurrent neural networks, and long short-term memory networks [11], [12], [13], [14], [15]. Although advanced models can learn complex representations, they generally require greater computational resources, larger datasets, and more extensive parameter optimization. In contrast, the K-Nearest Neighbor algorithm provides a simple and interpretable classification mechanism by assigning a class according to the proximity of feature vectors in the acoustic feature space. Machine learning algorithms have also demonstrated their applicability in broader classification and pattern analysis domains [16], [17], [18], [19]. Therefore, KNN remains a relevant lightweight baseline for evaluating the discriminative capability of MFCC features before more computationally intensive models are considered.

Another important consideration in speech analysis is the dataset's characteristics, particularly regarding speaker attributes [20], [21]. Many previous studies utilize datasets that combine speech recordings from multiple speakers with varying characteristics [11], [12], [13]. While such datasets provide diversity, they may also introduce additional variability that affects the consistency of feature distributions. In particular, differences in vocal properties between male and female speakers can influence acoustic patterns, potentially impacting classification performance.

Despite the progress reported in previous studies, several research limitations remain. First, many speech-related classification studies combine male and female recordings, although differences in pitch range, formant structure, vocal tract characteristics, and speaking style may introduce additional variability into the acoustic feature space. Second, recent studies increasingly emphasize complex deep learning architectures, while the effectiveness of lightweight and interpretable baseline classifiers for female-specific acoustic pattern classification has received comparatively less attention. Third, although MFCC features are widely used, limited discussion has addressed which predefined female voice categories remain acoustically separable and which exhibit substantial overlap when classified using a distance-based method. Consequently, an empirical evaluation focusing exclusively on female voice signals is required to assess the discriminative capability of MFCC features and to analyze class-specific misclassification patterns using a computationally simple classifier.

To provide a structured overview of the research context, a conceptual mind map illustrates the relationships among key components of acoustic speech analysis, including feature representation, classification methods, and dataset characteristics. The diagram highlights how variability in speech signals and the use of mixed-gender datasets contribute to the challenge of distinguishing acoustic patterns. More importantly, it emphasizes a research gap in the limited exploration of female-specific voice analysis and the relatively low adoption of simple baseline models such as KNN. Based on this gap, the study sets out to analyze acoustic patterns in female voice signals using MFCC features and a lightweight classification approach. The overall conceptual relationship between these elements is illustrated in Figure 1.

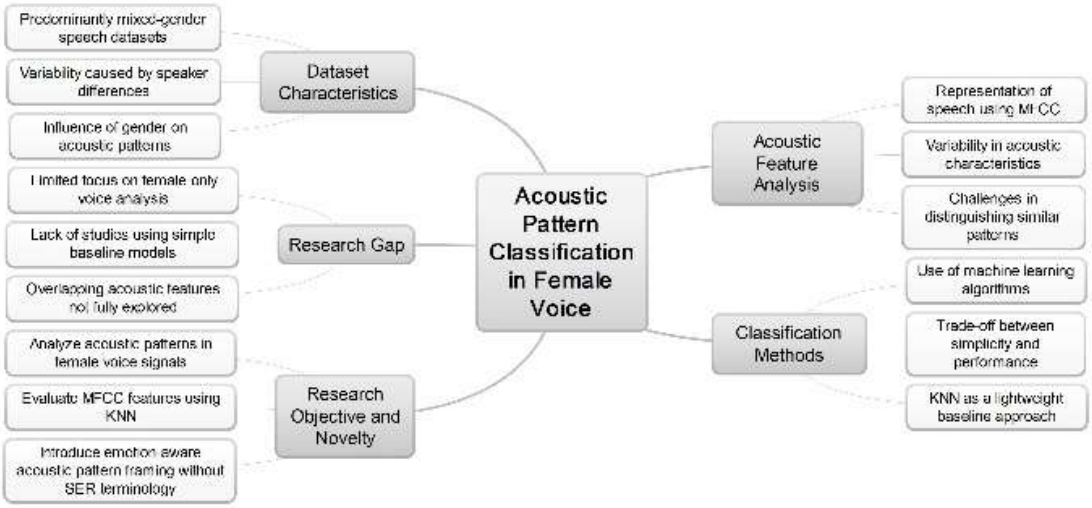

Figure 1. Conceptual framework of the research gap, objective, and contribution in female voice acoustic pattern classification

Figure 1 summarizes the research context formulated from the preceding literature review. It connects the main components of the study, including variability in speech signals, MFCC-based feature representation, classifier selection, and characteristics of the female-specific dataset. The figure does not independently define the research gap; rather, it visually reinforces the identified limitations concerning mixed-speaker variability, limited evaluation of lightweight classifiers, and insufficient class-level analysis of overlapping acoustic patterns. Based on these considerations, this study investigates the use of MFCC features and KNN for classifying predefined acoustic categories in female voice signals.

To address these limitations, this study focuses exclusively on female voice signals to reduce the variability associated with gender-related acoustic differences. Each speech sample is represented using a 58-dimensional MFCC feature vector, while the K-Nearest Neighbor algorithm is employed as a lightweight and interpretable classifier. The analysis covers predefined voice categories associated with the emotions angry, calm, disgusted, fearful, happy, neutral, sad, and surprised.

The objective of this study is to evaluate the effectiveness of MFCC features and KNN in classifying acoustic patterns in female voice signals. In addition to reporting overall classification performance, the study examines class-level precision, recall, F1-score, and confusion patterns to determine which categories have distinctive acoustic representations and which categories exhibit overlapping feature distributions.

The contribution of this study is a female-specific, computationally lightweight framework for acoustic pattern classification that uses a 58-dimensional MFCC representation and KNN. To the best of our knowledge, limited studies have explicitly examined both the overall effectiveness and class-level limitations of a lightweight MFCC–KNN framework using exclusively female voice samples. The study contributes by: (1) reducing gender-related variability through female-only data selection; (2) evaluating MFCC separability using an interpretable distance-based classifier; and (3) identifying category-specific acoustic overlap through detailed confusion matrix analysis. These findings provide a reproducible baseline for subsequent studies employing enhanced acoustic features or more advanced classification models.

II. METHODOLOGY

This study employed a supervised machine learning approach to classify acoustic patterns in female voice data using precomputed Mel-Frequency Cepstral Coefficient features and the K-Nearest Neighbor algorithm. The research procedure comprised dataset acquisition, female sample selection, data analysis, feature normalization, stratified data splitting, KNN model configuration, and performance evaluation. Because the source dataset already provided numerical MFCC values, this study operated at the feature level and did not perform direct audio recording or independent MFCC extraction from raw speech signals. The overall research procedure is illustrated in Figure 2.

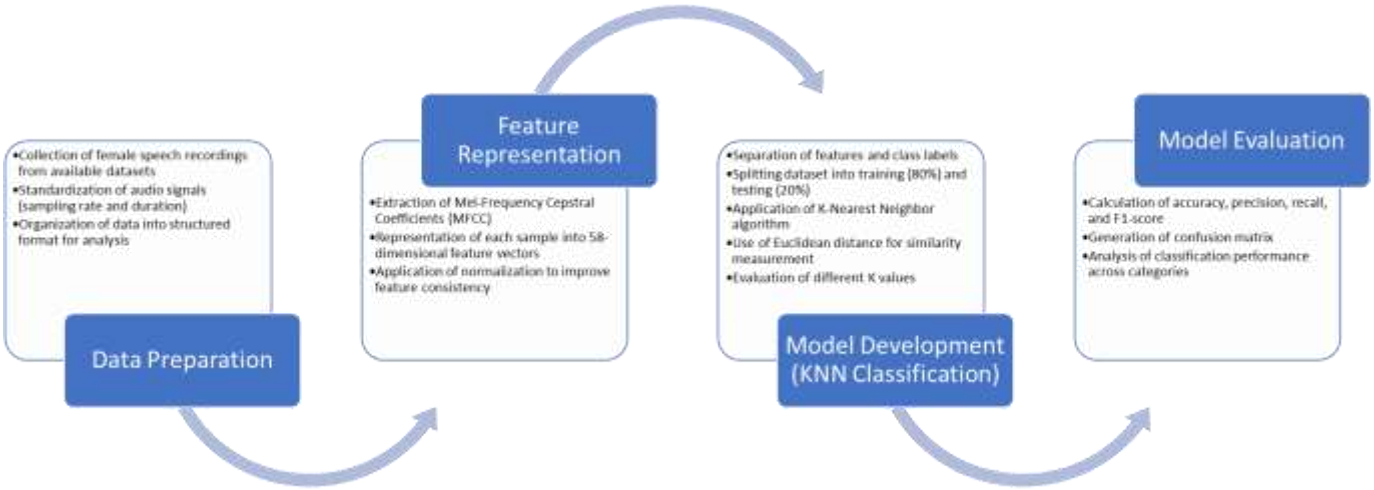

Figure 2. Research workflow for acoustic pattern classification using MFCC features and K-Nearest Neighbor

The dataset used in this study was obtained from the publicly available "MFCCs for Speech Emotion Recognition" dataset on Kaggle. The dataset can be accessed at https://www.kaggle.com/cracc97/emotional-speech-recognition and contains numerical MFCC-based representations of male and female voice samples [22], [23]. In this study, only female voice observations were selected in accordance with the research objective. After gender-based data selection, the dataset used in the experiment contained 49,225 observations distributed across eight predefined categories: angry, calm, disgust, fear, happy, neutral, sad, and surprise. Each observation consisted of 58 numerical MFCC features and one corresponding category label. The dataset was used in its feature-level form; therefore, this study did not independently obtain raw audio recordings or perform a new MFCC extraction.

The audio data are obtained from publicly available speech datasets containing recordings from multiple female speakers under varying recording conditions. These datasets provide a diverse range of vocal expressions, enabling the analysis of variations in acoustic characteristics across different speech samples. All audio recordings are processed to generate structured feature representations suitable for machine learning tasks. To ensure consistency across sources, the audio signals are standardized for sampling rate and duration before feature extraction. This step helps reduce variability caused by recording conditions and ensures that the extracted features are comparable across all samples.

Each observation in the source dataset was represented by 58 numerical features derived from the Mel-Frequency Cepstral Coefficients. The use of 58 dimensions was determined by the Kaggle dataset's original structure rather than by an independent comparison of different MFCC configurations. All 58 features were retained to preserve the dataset's complete acoustic representation and maintain consistency and reproducibility with the original data source.

The use of 58 dimensions should therefore not be interpreted as evidence that 58 coefficients are inherently superior to the commonly used configurations of 13, 20, or 40 coefficients. The number of MFCC features may vary depending on the coefficient configuration, dynamic features, and the aggregation strategy used during feature extraction [9], [10]. Because this study used precomputed feature-level data, reducing the representation to 13, 20, or 40 dimensions would require access to the original audio recordings and a new MFCC extraction experiment. No dimensionality reduction or manual feature selection was performed, and the complete set of 58 features was used as the input for KNN classification.

Each speech signal is represented using Mel-Frequency Cepstral Coefficients (MFCC), which capture important spectral properties of the audio signal based on a perceptually motivated frequency scale [24], [25]. In this study, each sample is encoded into a 58-dimensional MFCC feature vector. These features are commonly used in speech analysis due to their ability to represent the spectral envelope and preserve relevant acoustic information [26]. The dataset used in this research focuses exclusively on female voice recordings, with each instance corresponding to a feature vector and its associated class label from predefined acoustic categories [27]. In addition to MFCC extraction, basic normalization techniques are applied to the feature vectors to improve numerical stability and enhance the classification algorithm's performance.

TABLE I
DISTRIBUTION OF FEMALE VOICE DATASET

Category	Total Samples	Training Samples	Testing Samples
Angry	7,675	6,140	1,535
Calm	670	536	134
Disgust	7,670	6,136	1,534
Fear	7,675	6,140	1,535
Happy	7,675	6,140	1,535
Neutral	6,720	5,376	1,344
Sad	7,670	6,136	1,534
Surprise	3,470	2,776	694
Total	**49,225**	**39,380**	**9,845**

As shown in Table I, the female voice dataset contained 49,225 observations. A total of 39,380 observations were included in the training subset, while 9,845 observations were included in the testing subset. The distribution also shows that the calm and surprise categories contained fewer observations than the other categories. Therefore, class-specific, macro-average, and weighted-average metrics were considered in the performance analysis.

Before model development, an initial data examination is conducted to ensure the dataset's quality and consistency. This includes verifying the dimensionality of the extracted features, identifying missing or inconsistent values, and confirming the alignment between feature vectors and class labels. After this stage, the dataset is organized into two main components: input features and target labels, which are used for the classification process. Furthermore, an exploratory analysis is conducted to examine the distributions of feature values and class labels, providing an initial understanding of the dataset's characteristics before model training.

To evaluate the performance of the classification model, the dataset is divided into training and testing subsets. Approximately 80% of the data is allocated for training, while the remaining 20% are used for testing. A stratified sampling strategy is applied to maintain the proportional distribution of class labels in both subsets. The training data are used to construct the classification model, whereas the testing data are used to assess the model's generalization capability on unseen samples. This separation ensures that the evaluation results reflect the model's ability to handle new data rather than its ability to memorize the training samples.

The classification process is performed using the KNN algorithm. This distance-based learning method classifies data points based on their proximity to neighboring instances in the feature space [28], [29]. In this study, several K values are evaluated to determine the most suitable configuration for the classification task. The distance between feature vectors is calculated using the Euclidean distance metric, enabling the model to identify the nearest neighbors and assign class labels via majority voting. The simplicity of KNN makes it a suitable baseline for assessing the separability of acoustic features in the dataset. Additionally, the impact of different K values on classification performance is analyzed to identify the optimal neighborhood size.

To provide a direct comparative benchmark, a Support Vector Machine classifier was also evaluated using the same 58-dimensional MFCC representation, female voice dataset, training subset, and testing subset as the KNN model. SVM was selected because it represents a different classification mechanism: KNN assigns categories based on local neighborhood proximity, whereas SVM separates categories using decision boundaries in the feature space. Both models were evaluated using the same accuracy, precision, recall, F1-score, and training-time measurements to ensure a consistent comparison.

The performance of the proposed approach is assessed using standard evaluation metrics, including accuracy, precision, and recall. These metrics provide insight into the model's effectiveness in distinguishing between different acoustic patterns. Additionally, a confusion matrix is used to visualize classification outcomes and identify potential misclassification across categories. This evaluation process helps assess how well the KNN model captures variations in acoustic features in female voice signals. Further analysis is conducted by examining classification errors, which can provide useful information about overlapping feature distributions and potential limitations of the selected feature representation.

III. Result and Discussion

The performance of the proposed K-Nearest Neighbor (KNN) model in classifying acoustic patterns in female voice signals is evaluated using several standard metrics, including accuracy, precision, recall, and F1-score. As presented in Table 2, the model achieves an overall accuracy of 87.75%, with a weighted precision of 0.8792, recall of 0.8775, and F1-score of 0.8775. These results indicate that the selected MFCC features effectively represent acoustic characteristics, enabling the KNN model to distinguish among different voice patterns with relatively high reliability. In addition, the training time of approximately 0.10 seconds indicates a low model-fitting cost under the present experimental setting. However, because prediction latency and memory consumption were not measured, this result should not be interpreted as direct evidence of real-time deployment capability.

TABLE II
OVERALL PERFORMANCE OF KNN MODEL

Model	Precision	Accuracy	Recall	F1-Score	Training Time (s)
KNN	0.8792	0.8775	0.8775	0.8775	0.1036

A more detailed analysis of the classification results across different acoustic categories is presented in Table 3. The results show that the model performs differently across classes, depending on the distinctiveness of their acoustic features. The *surprise* category achieves the highest performance, with a precision of 0.9598, a recall of 0.9640, and an F1-score of 0.9619, indicating that this class has highly distinguishable acoustic patterns. Similarly, the *calm* category demonstrates strong recall (0.9552), suggesting that the model correctly identifies most samples in this category.

TABLE III
CLASSIFICATION PERFORMANCE ACROSS ACOUSTIC CATEGORIES

Category	Recall	Precision	F1-Score	Support
Angry	0.9231	0.8912	0.9069	1535
Calm	0.9552	0.8951	0.9242	134
Disgust	0.8801	0.8007	0.8385	1534
Fear	0.8886	0.8716	0.8800	1535
Happy	0.8176	0.9161	0.8640	1535
Neutral	0.8519	0.8774	0.8645	1344
Sad	0.8546	0.8805	0.8674	1534
Surprise	0.9640	0.9598	0.9619	694

The confusion matrix provides a detailed representation of the classification performance by illustrating the distribution of correct and incorrect predictions across all acoustic categories. The diagonal elements indicate correctly classified instances, while off-diagonal values reflect misclassification patterns between categories. This analysis provides a deeper understanding of how the KNN model distinguishes among acoustic patterns, as shown in Figure 3.

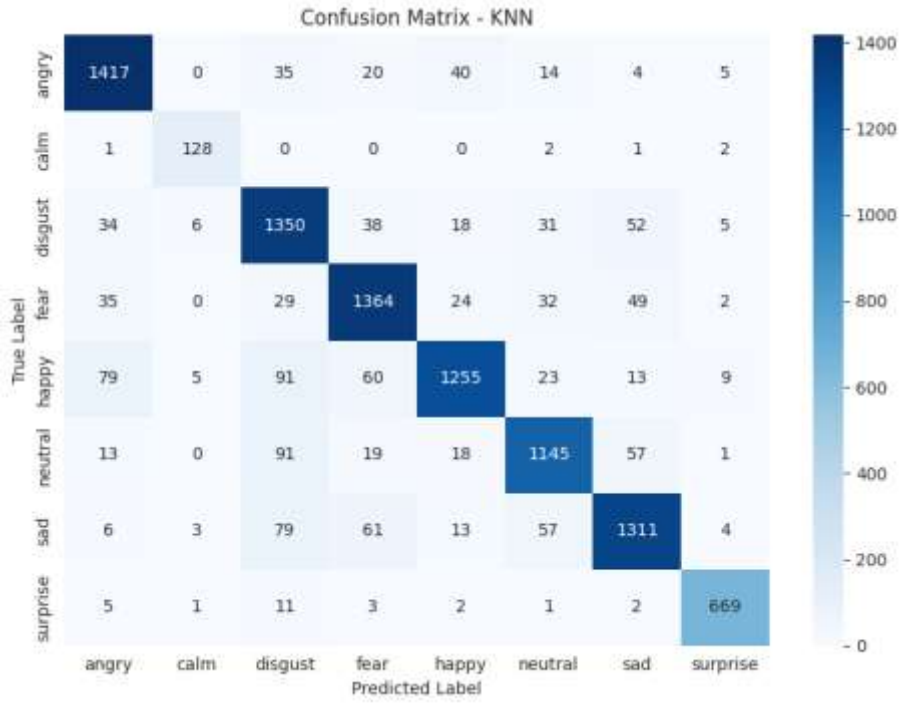

Figure 3. Confusion Matrix of KNN for Acoustic Pattern Classification in Female Voice Signals

The results show that several categories achieve a high number of correct predictions. The *angry* category records 1417 correctly classified samples out of 1535, while *fear* and *sad* also demonstrate strong performance with 1364 and 1311 correct

predictions, respectively. The *surprise* category exhibits one of the highest classification accuracies, with 669 correct predictions out of 694 samples, indicating that its acoustic characteristics are highly distinguishable. Similarly, the *calm* category achieves 128 correct predictions out of 134 samples, reflecting strong model performance despite its relatively smaller sample size.

However, notable misclassification patterns are observed in several categories. The *happy* category shows significant confusion, with 79 samples misclassified as angry, 91 as disgust, and 60 as fear, suggesting that these categories share overlapping acoustic features. In addition, the *neutral* category is frequently misclassified as *disgust* (91 samples) and *sad* (57 samples), indicating similarities in their spectral representations.

Further confusion is observed between *disgust* and *sad*, with 52 disgust samples predicted as sad, and between fear and sad, with 49 fear samples misclassified as sad. These patterns indicate that certain acoustic categories exhibit closely related feature distributions when represented using MFCCs, thereby reducing the effectiveness of distance-based classification.

Overall, the KNN results indicate that MFCC features provide a useful representation for classifying acoustic patterns of female voices. Nevertheless, the observed misclassification among acoustically similar categories demonstrates the limitations of relying exclusively on local distance in the feature space. To provide a clearer assessment of the relative performance of KNN, the model was subsequently compared with SVM using the same feature representation and data partition.

TABLE IV
OVERALL PERFORMANCE COMPARISON BETWEEN KNN AND SVM

Model	Precision	Accuracy	Recall	F1-Score	Training Time (s)
KNN	0.8792	0.8775	0.8775	0.8775	0.1036
SVM	0.8097	0.8085	0.8085	0.8082	128.1536

Table IV presents a direct comparison between KNN and SVM using the same female voice dataset, a 58-dimensional MFCC representation, and the same training–testing partition. KNN achieved an accuracy of 87.75%, whereas SVM achieved 80.85%. Thus, KNN outperformed SVM by approximately 6.90 percentage points in overall accuracy. KNN also achieved higher weighted precision, recall, and F1-score, with respective differences of approximately 6.95, 6.90, and 6.93 percentage points.

Under the present experimental configuration, KNN produced better classification performance than SVM, suggesting that local neighborhood relationships in the MFCC feature space were more suitable for the current dataset and data partition. The female voice categories may form locally concentrated groups that can be effectively recognized through proximity-based classification, while some category boundaries remain insufficiently distinct for SVM. However, this interpretation is limited to the current dataset and experimental partition and should not be generalized to all speech-emotion datasets.

At the category level, KNN obtained a higher F1-score than SVM in seven of the eight categories. The largest differences were observed for fear, disgust, and neutral. The F1-score for fear decreased from 0.8800 with KNN to 0.7818 with SVM, while disgust decreased from 0.8385 to 0.7475 and neutral decreased from 0.8645 to 0.7762. In contrast, SVM performed slightly better for surprise, achieving an F1-score of 0.9680 compared with 0.9619 for KNN. This result indicates that surprise had a strongly separable acoustic representation under both classifiers.

The reported training time also differed substantially. KNN required approximately 0.10 seconds, whereas SVM required approximately 128.15 seconds. This result shows that KNN required considerably less model-fitting time in the present experiment. However, prediction time was not measured; therefore, the reported training-time difference should not be interpreted as a complete comparison of end-to-end computational efficiency.

To further examine the category-level performance of the benchmark model, Figure 4 presents the confusion matrix generated by SVM. The diagonal values represent correctly classified observations, whereas the off-diagonal values indicate misclassification among the acoustic categories.

Figure 4. Confusion Matrix of SVM for Acoustic Pattern Classification in Female Voice Signals

As shown in Figure 4, SVM correctly classified 1,341 angry, 128 calm, 1,168 disgust, 1,143 fear, 1,180 happy, 1,042 neutral, 1,278 sad, and 680 surprise observations. The surprise category achieved the highest recall, followed by calm and angry. However, there was greater confusion in the disgust, fear, happy, neutral, and sad categories.

Among the happy observations, 96 were classified as angry, 90 as fear, and 83 as disgust. For disgust, 108 observations were classified as sad, 98 as neutral, 62 as angry, and 49 as happy. Fear was also frequently classified as sad, with 157 observations assigned to that category. These results indicate that the overlap was not limited to happy and disgust but extended to several categories with partially similar MFCC distributions.

The confusion matrices for both KNN and SVM indicate that the observed classification errors were due to partially overlapping acoustic representations across several categories. Emotion-related vocal expressions are communicated through combinations of acoustic cues, including pitch, intensity, timing, speech rate, vocal quality, and spectral characteristics, rather than through a single acoustic attribute. Consequently, variations in emotional intensity and speaking style may cause observations from different categories to occupy adjacent regions of the feature space. Because the present dataset uses precomputed MFCC features, longer-term prosodic changes and temporal trajectories may not be captured as completely as short-term spectral characteristics. This limitation provides a plausible explanation for the repeated confusion among happy, disgust, angry, fear, neutral, and sad.

The class imbalance shown in Table 1 should also be considered when interpreting the results. The calm category contained only 134 testing observations, compared with more than 1,500 observations in several other categories. Both KNN and SVM correctly classified 128 of the 134 calm samples, resulting in the same recall of 0.9552. However, KNN achieved a calm precision of 0.8951, whereas SVM obtained 0.7853. This difference indicates that SVM assigned more observations from other categories to calm, despite correctly identifying most actual calm observations.

Because calm contained only 134 testing samples, each additional error would change its recall by approximately 0.75 percentage points. Therefore, its high recall should be interpreted with caution, alongside precision, F1-score, and sample support. Detailed class-level metrics are particularly important in imbalanced multi-class emotion-recognition problems because overall accuracy alone may conceal weaker performance in underrepresented categories.

The evaluation was based on a single training–testing partition. Confidence intervals and repeated or k-fold cross-validation were not calculated. Therefore, the reported results describe performance on the current testing subset but do not quantify variation across alternative data partitions. This limitation should be considered when interpreting the stability and generalizability of both models.

The comparative findings position KNN as the better-performing classifier under the present experimental setting. KNN achieved higher overall accuracy, weighted precision, recall, and F1-score than SVM and required substantially less time to fit the model. Nevertheless, the confusion matrices of both models reveal persistent classification challenges among acoustically overlapping categories. The results should therefore be interpreted within the context of the unequal category distribution, the precomputed MFCC representation, and the use of a single training–testing partition.

IV. CONCLUSION

This study evaluated the use of a 58-dimensional MFCC representation and the K-Nearest Neighbor algorithm for classifying acoustic patterns in female voice signals. The experimental results showed that KNN achieved an accuracy of 87.75%, a weighted precision of 0.8792, a recall of 0.8775, and an F1-score of 0.8775. In a direct comparison using the same dataset and data partition, KNN outperformed SVM, achieving an accuracy of 80.85% and an F1-score of 0.8082. KNN also obtained higher class-level F1-scores in seven of the eight categories and required substantially less model-fitting time under the present experimental configuration.

The category-level results showed that surprise and calm were recognized with relatively high recall, whereas happy, disgust, fear, neutral, and sad exhibited greater classification overlap. The confusion matrices indicate that MFCC features can represent several distinguishable acoustic patterns, but some categories remain difficult to separate because their spectral characteristics occupy adjacent regions of the feature space. The results also showed that the high recall for the calm category should be interpreted with caution because it was calculated from only 134 test observations, considerably fewer than the support for most other categories.

In practical terms, the proposed MFCC–KNN framework can serve as a lightweight baseline for analyzing female-voice acoustic patterns in applications such as affect-aware human–computer interaction, voice-based educational systems, call-center speech analytics, and preliminary emotion-related voice screening. However, the present results should not be interpreted as evidence of production-ready or real-time deployment, as prediction latency, memory consumption, and performance under live-recording conditions were not evaluated.

This study has several limitations. The analysis was restricted to female voice data and used precomputed 58-dimensional MFCC features, which may not fully represent longer-term prosodic and temporal characteristics. The category distribution was unequal, particularly for calm and surprise. In addition, the evaluation was based on a single training–testing partition without confidence intervals, repeated cross-validation, or systematic optimization of the KNN neighborhood parameter. Therefore, the reported results describe performance under the current experimental configuration and may vary under different data partitions or parameter settings.

Future research should evaluate repeated stratified k-fold cross-validation and systematic hyperparameter optimization to obtain more stable performance estimates. Class-weighting, oversampling, or other imbalance-handling techniques should also be investigated. More advanced models, including convolutional neural networks applied to spectrograms, LSTM or GRU networks for modeling temporal acoustic dynamics, and Transformer-based architectures for capturing longer-range dependencies, may improve the recognition of acoustically overlapping categories. Combining MFCC with prosodic features such as pitch, energy, speaking rate, and temporal contours is also recommended to provide a more comprehensive representation of female voice patterns.

ACKNOWLEDGMENT

The author expresses sincere appreciation to Prof. Sofyan Anif, former Rector of Universitas Muhammadiyah Surakarta, for his distinguished leadership and valuable contributions during his tenure. The author also gratefully acknowledges Prof. Harun Joko Prayitno, current Rector of Universitas Muhammadiyah Surakarta, for his ongoing support and commitment to advancing academic research and institutional development.

REFERENCES

[1] S. Gayathri, J. P. Kadambarajan, S. Harita, T. Rakshani, and K. Gayathri, “Identification of Voice Disorderliness Using Machine Learning Algorithm,” in *2025 International Conference on Computing and Communication Technologies, ICCCT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICCCT63501.2025.11019856.

[2] Y. Xia, Y. He, B. Kang, Y. Wang, X. Dong, and Z. Zhang, “Research and application of MFCC-LSTM based on improved ZOA algorithm to optimize transformer voice-print diagnosis,” in *2024 4th International Conference on Electrical Engineering and Control Science, IC2ECS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 559–563. doi: 10.1109/IC2ECS64405.2024.10928641.

[3] L. Zhang *et al.*, “Voiceprint Unlocking Based on MFCC - Exploration of Voiceprint Models Different from Fingerprint,” in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications, ICIPCA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 763–769. doi: 10.1109/ICIPCA61593.2024.10709042.

[4] R. Ranjan, M. K. Singh, and A. Sharma, “Text Dependent Speaker Identification from Disguised Voice Using Feature Extraction and Classification,” in *Proceedings of International Conference on Emerging Technologies and Innovation for Sustainability, EmergIN 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 181–186. doi: 10.1109/EmergIN63207.2024.10961932.

[5] J. Hong, J. Lee, D. Cho, and J. Jung, “Depression Classification Algorithm Based on Voice Signals Using MFCC and CNN Autoencoders,” in *Proceedings - 2024 International Conference on Machine Learning and Applications, ICMLA*

2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1368–1373. doi: 10.1109/ICMLA61862.2024.00213.

[6] A. Chakhtouna, S. Sekkate, and A. Adib, "Speech emotion recognition: A systematic mega-review of techniques and pipelines," *Information Fusion*, vol. 131, p. 104161, 2026, doi: https://doi.org/10.1016/j.inffus.2026.104161.

[7] S. A. Jakhete and N. Kulkarni, "Enhanced Human Emotion Recognition through Multimodal Data using Deep Learning and Late Fusion Technique," *International Journal of Engineering*, vol. 39, no. 9, pp. 2177–2188, 2026, doi: 10.5829/ije.2026.39.09c.09.

[8] G. P. Chaitra and M. Farida Begam, "Real Time Dialectal Speech Recognition Using MFCCs on Mobile Applications," in *Proceedings of 2025 International Conference on Emerging Technologies in Computing and Communication, ETCC 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ETCC65847.2025.11108382.

[9] B. Raviteja, N. Adithi, P. S. Jashwanth, and L. Sunitha, "Affective Speech Processing Using MFCC," in *2025 International Conference on New Trends in Computing Sciences, ICTCS 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 239–244. doi: 10.1109/ICTCS65341.2025.10989400.

[10] M. Sivaramakrishnan, A. Rajput, and M. Saravanan, "Classification of Deep Fake Audio Using MFCC Technique," in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems, ICITEICS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICITEICS61368.2024.10625305.

[11] S. Nath, A. K. Shahi, T. Martin, N. Choudhury, and R. Mandal, "Speech Emotion Recognition Using Machine Learning: A Comparative Analysis," *SN Comput. Sci.*, vol. 5, no. 4, p. 390, Apr. 2024, doi: 10.1007/s42979-024-02656-0.

[12] K. Nandini, T. Divya, S. Subhani, S. Mounika, and T. Vamsi Nandhan, "Recognition of emotional speech using MFCC and Machine Learning Technique," in *ESIC 2024 - 4th International Conference on Emerging Systems and Intelligent Computing, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 416–421. doi: 10.1109/ESIC60604.2024.10481531.

[13] A. Benzirar, M. Hamidi, and M. Filali Bouami, "Building a speech emotion recognition system using RNN, GRU and LSTM," *Int. J. Speech Technol.*, vol. 28, no. 3, pp. 745–759, Sep. 2025, doi: 10.1007/s10772-025-10214-z.

[14] I. Puspitasari, A. Yudhana, D. E. Wati, and S. Al Irfan, "Recognizing Micro Expression Pattern Using Convolutional Neural Networks (CNN) Method During Emotion Regulation Training for Parents in The Pandemic Era," 2023. [Online]. Available: https://icistech.org/index.php/icistech/

[15] L. Li, X. Chen, and P. Zhu, "How do e-commerce anchors' characteristics influence consumers' impulse buying? An emotional contagion perspective," *Journal of Retailing and Consumer Services*, vol. 76, p. 103587, 2024, doi: https://doi.org/10.1016/j.jretconser.2023.103587.

[16] K. Zahra Nurbana and E. Sudarmilah, "Non-fungible token modeling: the enthusiasm of music fans for the digital-collectible revolution," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 4, pp. 2880–2888, Aug. 2025, doi: 10.11591/eei.v14i4.9269.

[17] Y. I. Kurniawan, F. Razi, N. Nofiyati, B. Wijayanto, and M. L. Hidayat, "Naive Bayes modification for intrusion detection system classification with zero probability," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2751–2758, Oct. 2021, doi: 10.11591/eei.v10i5.2833.

[18] A. G. Putrada, N. Alamsyah, S. F. Pane, M. N. Fauzan, and D. Perdana, "Virtual Sensors Method and Architecture for a Smart Home Environment with Random Forest," in *2023 10th International Conference on ICT for Smart Society (ICISS)*, IEEE, Sep. 2023, pp. 1–6. doi: 10.1109/ICISS59129.2023.10292065.

[19] A. G. Putrada, N. Alamsyah, M. N. Fauzan, and D. Perdana, "PCA-SVM for a Lightweight ASL Hand Gesture Image Recognition," in *Proceedings of the International Conference on Electrical Engineering and Informatics*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICEEI59426.2023.10346744.

[20] M. Adeel, Z.-Y. Tao, S.-Y. Jin, C.-J. Guo, and M. Alsuhaibani, "Assessing random forest performance in low resource speech emotion recognition," *Sci. Rep.*, vol. 16, no. 1, p. 854, Dec. 2025, doi: 10.1038/s41598-025-30511-6.

[21] J. Mi, X. Shi, D. Ma, J. He, T. Fujimura, and T. Toda, "Robust speech emotion recognition under human speech noise," *Comput. Speech Lang.*, vol. 100, p. 101987, Oct. 2026, doi: 10.1016/j.csl.2026.101987.

[22] D. P. Major and M. Chatterjee, "Acoustic analyses of the RAVDESS corpus of emotional stimuli," *JASA Express Lett.*, vol. 6, no. 2, Feb. 2026, doi: 10.1121/10.0042364.

[23] Z. S. Kahhoul, N. Terki, M. L. Tiar, I. Benaissa, and S. Boutiba, "Ensemble Learning for Improved Speech Emotion Recognition: A Control Dimension Analysis of Log-Mel Spectrograms from the CREMA-D Dataset," *Signal Image Video Process.*, vol. 19, no. 13, p. 1135, Dec. 2025, doi: 10.1007/s11760-025-04763-8.

[24] R. Nutenki, A. Thatipudi, A. K. Perikala, and H. Medida, "Analysis and Classification of Arcing Signals by Using MFCC," in *2024 4th International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICAECT60202.2024.10468692.

[25] V. Rraci, L. Anderson, and S. Chakrabarty, “Classification of Right Hemisphere Damage Using MFCC Paralinguistic Voice Features,” in *2025 Intermountain Engineering, Technology and Computing, IETC 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/IETC64455.2025.11039385.

[26] W. Liu and Y. Duan, “Speaker recognition based on MFCC and GFCC feature parameter extraction,” in *2025 5th International Symposium on Computer Technology and Information Science, ISCTIS 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 33–37. doi: 10.1109/ISCTIS65944.2025.11065212.

[27] M. Tummala, L. Harish, M. E. Malkhed, S. S. Kumar, N. Neelima, and V. Venugopal, “Optimizing Gender Identification with MFCC Feature Engineering and Enhanced Gradient Boosting,” in *2024 Asian Conference on Intelligent Technologies, ACOIT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ACOIT62457.2024.10939536.

[28] Y. Bai, “A Study on Speech Emotion Recognition Based on MFCC and KNN Models,” in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications, ICIPCA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 729–732. doi: 10.1109/ICIPCA61593.2024.10709039.

[29] Md. N. Alam Siddiquee, Md. A. Hossain, and F. Wahida, “An Effective Machine Learning Approach for Music Genre Classification with Mel Spectrograms and KNN,” in *2023 International Conference on Communication, Circuits, and Systems (IC3S)*, IEEE, May 2023, pp. 1–4. doi: 10.1109/IC3S57698.2023.10169397.