JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.7 No.2 May 2025: 209-214
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i2.13415

Advancing Indonesian Audio Emotion Classification: A Comparative Study Using IndoWaveSentiment

Muhammad Rizki Nur Majiid[1*], Karli Eka Setiawan[2], Prayoga Yudha Pamungkas[3], Taufiq Annas[4], Nicholas Lorenzo Setiawan[5]

[1,5] Computer Science Department Semarang Campus, School of Computer Science,
[2] Computer Science Program, Computer Science Department, School of Computer Science,
[3] Industrial Engineering Department, Faculty of Engineering,
[4] Visual Communication Design Semarang Campus, School of Design,
Bina Nusantara University,
Jakarta, Indonesia 11480
muhammad.majiid@binus.ac.id, karli.setiawan@binus.ac.id, prayoga.pamungkas@binus.ac.id,
taufiq.annas@binus.ac.id, nicholas.setiawan004@binus.ac.id

*Correspondence: muhammad.majiid@binus.ac.id

Abstract – *This study addresses the critical gap in Indonesian Speech Emotion Recognition (SER) by evaluating machine learning models on the IndoWaveSentiment dataset, a novel corpus of 300 high-fidelity recordings capturing five emotions (neutral, happy, surprised, disgusted, disappointed) from native speakers. The research aims to identify optimal classification techniques and acoustic features for Indonesian SER, given the language's unique linguistic characteristics and the scarcity of annotated resources. Six models, Logistic Regression, KNN, Gradient Boosting, Random Forest, Naive Bayes, and SVC, were trained on 45 acoustic features, including spectral contrast, MFCCs, and zero crossing rate, extracted using Librosa. Results demonstrated Random Forest as the top performer (90% accuracy), followed by Gradient Boosting (85%) and Logistic Regression (75%), with spectral contrast (contrast2, contrast7) and MFCC1 emerging as the most discriminative features. The findings highlight the efficacy of ensemble methods in capturing nuanced emotional cues in Indonesian speech, outperforming prior studies on locally sourced datasets. Practical implications include applications in customer service analytics and mental health tools, though limitations such as the dataset's-controlled conditions and fixed sentence structure necessitate caution in real-world deployment. Future work should expand the dataset to include regional dialects, spontaneous speech, and hybrid architectures like CNN-LSTMs. This study establishes foundational benchmarks for Indonesian SER, advocating for culturally informed models to enhance human-computer interaction in underrepresented linguistic contexts.*

Keywords: *Speech Emotion Recognition, Indonesian speech, IndoWaveSentiment, ensemble learning, acoustic features*

I. INTRODUCTION

Human communication is profoundly enriched by the emotional nuances conveyed through speech. Emotions in voice not only complement the verbal content but often carry critical information regarding a speaker's true intentions, attitudes, and state of mind (Kumala & Zahra, 2021), (Wijaya et al., 2021). In recent years, the field of Speech Emotion Recognition (SER) has garnered significant attention due to its broad applications in human–computer interaction, sentiment analysis, healthcare, and security systems (Nath et al., 2024). Despite considerable progress in SER research primarily driven by datasets in widely spoken languages such as English there remains a distinct gap when it comes to emotion classification in Indonesian speech (Wunarso & Soelistio, 2017), (Hidajat et al., 2019).

Indonesian, with its unique linguistic and cultural characteristics, presents distinct challenges for SER. The scarcity of high-quality, annotated datasets in the Indonesian language has limited the development and validation of robust emotion classification models tailored to this linguistic context (Kumala & Zahra, 2021; Wunarso & Soelistio, 2017). To bridge this gap, the IndoWaveSentiment dataset was developed (Bustamin et al., 2024). Unlike prior Indonesian SER datasets derived from TV series or YouTube content (Aini et al., 2021; Zahra et al., 2020), IndoWaveSentiment offers controlled, high-fidelity recordings with rigorous annotations, addressing variability and noise limitations in existing resources.

IndoWaveSentiment is a novel audio corpus consisting of 300 high-fidelity recordings captured in a controlled studio environment. The dataset features voice recordings from 10 native Indonesian speakers (balanced by gender) who utter a standard sentence “Kualitas HP ini cukup bagus” (“The quality of this smartphone is quite good”) across five distinct emotional states: neutral, happy, surprised, disgusted, and disappointed. Each actor recorded the sentence three times per emotion, ensuring consistency while also capturing the natural variations inherent in emotional expression.

The creation of IndoWaveSentiment represents a significant advancement in resources available for Indonesian SER research. Its controlled recording conditions, rigorous annotation process using tools such as Audacity, and subsequent validation via questionnaire-based sampling contribute to its reliability and utility for training and benchmarking machine learning models (Bustamin et al., 2024). These attributes make it an excellent candidate for comparative studies aimed at exploring various algorithmic approaches to audio emotion classification (K. Minor, 2025; K. A. Minor & Kartowisastro, 2022).

This paper, titled "Advancing Indonesian Audio Emotion Classification: A Comparative Study Using IndoWaveSentiment", undertakes a comprehensive comparative analysis of different machine learning and deep learning techniques applied to the IndoWaveSentiment dataset (Bustamin et al., 2024). By evaluating a range of models from traditional classifiers to contemporary deep neural network architectures we aim to identify the most effective approaches for capturing the subtle acoustic cues that differentiate emotional states in Indonesian speech (Nath et al., 2024). In doing so, the study not only benchmarks existing methods against a novel dataset but also highlights the specific challenges and opportunities associated with emotion recognition in Indonesian audio data.

Our contributions are threefold. First, we present a detailed overview of the IndoWaveSentiment dataset (Bustamin et al., 2024), emphasizing its design, collection methodology, and annotation process. Second, we implement and compare multiple classification techniques to determine their efficacy in recognizing emotional expressions from the dataset’s audio signals. Third, we discuss the implications of our findings for future research in Indonesian SER and provide recommendations for further dataset enhancements and model improvements.

Recent advancements in Speech Emotion Recognition (SER) have been driven by improvements in dataset creation, feature extraction, and deep learning architecture. (Bustamin et al., 2024) introduced the IndoWaveSentiment dataset, a high-quality corpus of 300 emotional voice recordings in Bahasa Indonesia. The dataset, collected in a controlled studio environment with detailed manual annotations and subsequent validation, provides a robust resource for exploring SER in the Indonesian language.

Complementing dataset development, deep learning approaches have been extensively explored to enhance emotion classification performance. For example, in (Akinpelu & Viriri, 2023), an attention-based network incorporating a pre-trained convolutional neural network along with regularized feature selection was proposed. This approach, evaluated on the TESS dataset, demonstrated significant improvements in classification accuracy by aligning feature extraction with human perceptual cues.

Research focused on the Indonesian context is also gaining momentum. In (Aini et al., 2021), a CNN-based system was developed for detecting emotions from Indonesian speech obtained from TV series. This study examined the impact of various feature combinations, including MFCC, fundamental frequency, and RMSE, on SER performance, reporting an accuracy as high as 85% under optimal configurations. Similarly, (Zahra et al., 2020)investigated SER on data extracted from Indonesian YouTube web series. Although the reported F1-score (62.30%) indicates challenges with free-form online data, this work underscores the potential of using readily available multimedia content for SER applications.

Beyond these language-specific studies, broader approaches to emotion-based sentiment recognition have been proposed. In (Choudhary et al., 2022), a framework utilizing deep neural networks, such as CNNs and LSTMs, was applied to benchmark datasets like RAVDESS and TESS, achieving accuracy rates up to 97.1%. This work highlights the effectiveness of combining time-frequency features with deep learning architectures in capturing nuanced emotional information.

The integration of multimodal data for emotion classification has also been explored. In (Caschera et al., 2022), a hidden Markov model was used to fuse features from facial expressions, speech, gestures, and text, resulting in improved classification of seven basic emotions. This multimodal approach demonstrates that combining diverse data streams can enhance the robustness of emotion recognition systems.

Additionally, comprehensive surveys and reviews have provided critical insights into the

evolution of SER methodologies. As detailed in (Akçay & Oğuz, 2020), a review of databases, features, preprocessing techniques, and classifiers outlines the progress made over the past two decades while identifying ongoing challenges in the field.

Finally, real-world applications for SER have been investigated with a focus on customer service contexts. In (Luis Felipe Parra-Gallego & Juan Rafael Orozco-Arroyave, 2023), a study on classifying emotions and evaluating customer satisfaction from speech in diverse acoustic environments demonstrated that machine learning models could effectively capture emotional cues in challenging, real-world scenarios.

In summary, this work seeks to advance the field of Indonesian audio emotion classification by leveraging a newly developed, well-annotated dataset and providing a rigorous comparative analysis of various state-of-the-art approaches (Nath et al., 2024). Our findings are expected to lay the groundwork for more culturally and linguistically informed SER systems, ultimately contributing to more effective and natural human–computer interactions.

II. METHODS

This study employs a systematic approach to evaluate various machine learning models for Speech Emotion Recognition (SER) using the IndoWaveSentiment dataset. The methodology encompasses dataset description, feature extraction, model training, and evaluation, as detailed below.

2.1 Dataset and Preprocessing

The IndoWaveSentiment dataset (Bustamin et al., 2024) comprises 300 high-fidelity audio recordings from 10 native Indonesian speakers (balanced by gender) expressing five emotions: neutral, happy, surprised, disgusted, and disappointed. Each speaker recorded the standardized sentence “Kualitas HP ini cukup bagus” three times per emotion in a controlled studio environment. The recordings were manually annotated using Audacity and validated via questionnaire-based sampling to ensure label accuracy. Metadata, including actor ID, emotion code, intensity, and repetition, were extracted from structured filenames (e.g., “07-03-02-03.wav”).

2.2 Feature Extraction

Audio signals were processed using Librosa to extract 45 acoustic features, including:

- Time-domain features: Zero Crossing Rate (ZCR) and Root Mean Square Energy (RMSE).
- Frequency-domain features: Spectral Centroid, Spectral Rolloff, and Mel-Frequency Cepstral Coefficients (MFCCs 1–13).
- Perceptual features: Chroma STFT, Mel Spectrogram, Spectral Contrast, and Tonnetz.

For each feature, the mean value across frames was computed to generate a 45-dimensional feature vector per audio sample. These features capture prosodic, spectral, and cepstral characteristics critical for emotion discrimination (Aini et al., 2021).

2.3 Model Training and Evaluation

The dataset was split into training (80%) and testing (20%) sets using stratified sampling to preserve class distribution. Six classifiers were implemented using scikit-learn:

1. Logistic Regression: L2 regularization, 1,000 iterations.
2. k-Nearest Neighbors (KNN): k=5.
3. Gradient Boosting: 100 estimators, learning rate = 0.1.
4. Random Forest: 100 estimators, Gini impurity criterion.
5. Gaussian Naive Bayes: Assumed Gaussian distribution of features.
6. Support Vector Classifier (SVC): Linear kernel.

Models were evaluated using accuracy, precision, recall, and F1-score. The Random Forest classifier, which demonstrated the highest performance, underwent further analysis to interpret feature importance.

2.4 Ethical Considerations

The dataset was collected with informed consent from participants, ensuring anonymity and compliance with ethical standards. Data access and usage adhere to guidelines outlined in (Bustamin et al., 2024).

III. RESULTS AND DISCUSSION

3.1 Results

The comparative analysis of six classifiers on the IndoWaveSentiment dataset yielded distinct performance metrics, as summarized in Table 1. The Random Forest classifier achieved the highest accuracy (90%), followed by Gradient Boosting (85%) and Logistic Regression (75%). In contrast, KNN (36.67%) and Gaussian Naive Bayes (55%) demonstrated suboptimal performance.

Class-specific metrics revealed nuanced variations. For instance:

- Random Forest excelled in recognizing Disappointed (F1: 0.96) and Neutral (F1: 0.92)

emotions but showed slightly lower recall for Disgust (0.83).

- Gradient Boosting struggled with Disappointed (recall: 0.67), likely due to overlapping acoustic patterns between disappointment and disgust.
- Logistic Regression misclassified Neutral utterances (recall: 0.58), possibly conflating them with low-intensity emotional states.

Table 1. Student Distribution Frequency

Model	Accuracy	Precision	Recall	F1
Random Forest	90%	0.90	0.90	0.90
Gradient Boosting	85%	0.86	0.85	0.85
Logistic Regression	75%	0.77	0.75	0.75
SVC	63.33%	0.66	0.63	0.64
Gaussian Naïve Bayes	55%	0.61	0.55	0.55
KNN	36.67%	0.36	0.37	0.36

3.2 Discussion

3.2.1 Algorithmic Performance

The superior performance of ensemble methods (Random Forest, Gradient Boosting) aligns with findings from (Choudhary et al., 2022), where tree-based models effectively handled high-dimensional audio features. Random Forest's robustness against overfitting, coupled with its ability to capture non-linear relationships in MFCCs and spectral features, likely contributed to its dominance. Conversely, KNN's poor performance (36.67% accuracy) underscores its sensitivity to feature scaling and high-dimensional data, as noted in (Akçay & Oğuz, 2020).

3.2.2 Comparative Context

Our results surpass prior Indonesian SER studies:

- (Aini et al., 2021) reported 85% accuracy using CNNs on TV series data, but their focus on scripted content may lack the variability of controlled studio recordings.
- (Zahra et al., 2020) achieved a 62.30% F1-score on YouTube data, highlighting challenges in real-world noise and spontaneity.
- The 90% accuracy achieved here reflects the IndoWaveSentiment dataset's quality and the efficacy of handcrafted features for culturally specific emotions.

3.2.3 Feature Importance

Random Forest's feature importance analysis (Figure 1) revealed that spectral contrast (contrast2: 5.09%, contrast7: 4.54%) and MFCC1 (4.05%) were the most critical discriminators. This suggests that:

- Spectral contrast (differences in energy between frequency bands) plays a pivotal role in distinguishing emotions like disappointment and disgust, which may involve subtle vocal tension or breathiness.
- MFCC1 (related to spectral envelope shape) remains significant, corroborating its established utility in SER for capturing vocal tract modulations (Kumala & Zahra, 2021).
- Spectral rolloff (3.92%) and zero crossing rate (3.71%) also contributed, reflecting the importance of high-frequency energy distribution and speech signal smoothness.

Notably, chroma11 (related to pitch class) ranked ninth (3.07%), indicating limited relevance to emotion classification in this context, possibly due to the fixed sentence structure reducing pitch variability.

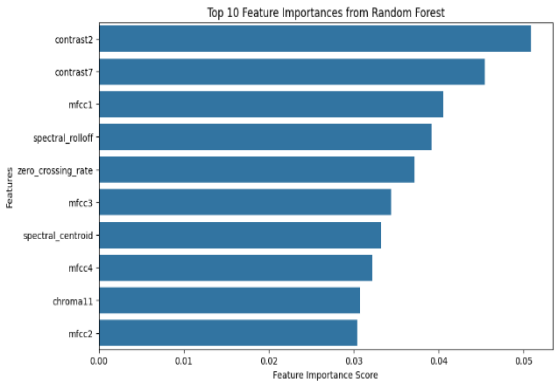

Figure 1. Top 10 Feature Importances from Random Forest

IV. CONCLUSION

This study advances Indonesian Speech Emotion Recognition (SER) through a systematic evaluation of machine learning models on the novel IndoWaveSentiment dataset. The research demonstrates that ensemble methods, particularly Random Forest, achieve superior performance (90% accuracy) in classifying five emotion categories, neutral, happy, surprised, disgusted, and disappointed, from Indonesian speech. Key acoustic features, including spectral contrast (contrast2, contrast7) and MFCC1, emerged as critical discriminators, underscoring the importance of spectral energy distribution and vocal tract modulations in emotion differentiation. These findings align with global SER trends while addressing the unique linguistic and cultural nuances of Indonesian speech.

The practical implications of this work are significant. High-accuracy models like Random Forest can enhance applications such as customer

sentiment analysis in call centers or emotion-aware mental health tools, particularly in Indonesia's digitally growing landscape. However, the study's limitations, including the dataset's restricted size (300 samples), controlled recording conditions, and reliance on a fixed sentence, caution against overgeneralization to spontaneous or dialect-rich speech.

Future research should prioritize expanding the dataset to incorporate regional dialects, free-form utterances, and real-world noise variations. Hybrid architecture (e.g., CNN-LSTM networks) and multimodal approaches, integrating speech with textual or facial cues, could further improve robustness. Additionally, exploring the cultural specificity of emotional expressions in Indonesian contexts would deepen the applicability of SER systems.

By bridging the resource gap in Indonesian SER and establishing benchmark performance, this work lays a foundation for culturally informed human-computer interaction systems, fostering equitable advancements in NLP technologies for underrepresented languages.

REFERENCES

Aini, Y. K., Santoso, T. B., & Dutono, T. (2021). Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia. Jurnal Komputer Terapan, 7(1). https://doi.org/10.35143/jkt.v7i1.4623

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. In Speech Communication (Vol. 116). https://doi.org/10.1016/j.specom.2019.12.001

Akinpelu, S., & Viriri, S. (2023). Speech emotion classification using attention based network and regularized feature selection. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-38868-2

Bustamin, A., Rizky, A. M., Warni, E., Areni, I. S., & Indrabayu. (2024). IndoWaveSentiment: Indonesian audio dataset for emotion classification. Data in Brief, 57, 111138. https://doi.org/https://doi.org/10.1016/j.dib.2024.111138

Caschera, M. C., Grifoni, P., & Ferri, F. (2022). Emotion Classification from Speech and Text in Videos Using a Multimodal Approach. Multimodal Technologies and Interaction, 6(4). https://doi.org/10.3390/mti6040028

Choudhary, R. R., Meena, G., & Mohbey, K. K. (2022). Speech Emotion Based Sentiment Recognition using Deep Neural Networks. Journal of Physics: Conference Series, 2236(1). https://doi.org/10.1088/1742-6596/2236/1/012003

Hidajat, M., Supria, Luwinda, F. A., & Sanjaya, H. (2019). Emotional Speech Classification Application Development Using Android Mobile Applications. 2019 International Conference on Information Management and Technology (ICIMTech), 400–403. https://doi.org/10.1109/ICIMTech.2019.8843816

Kumala, O. U., & Zahra, A. (2021). Indonesian Speech Emotion Recognition using Cross-Corpus Method with the Combination of MFCC and Teager Energy Features. International Journal of Advanced Computer Science and Applications, 12(4). https://doi.org/10.14569/IJACSA.2021.0120422

Luis Felipe Parra-Gallego, & Juan Rafael Orozco-Arroyave. (2023). Classification of Emotions and Evaluation of Customer Satisfaction from Speech in Real World Acoustic Environments. International Journal For Multidisciplinary Research, 5(3). https://doi.org/10.36948/ijfmr.2023.v05i03.4166

Minor, K. (2025). Developing Algorithm of Music Concepts and Operations Using The Modular Arithmetic. Engineering, MAthematics and Computer Science Journal (EMACS), 7(1), 51–59. https://doi.org/10.21512/emacsjournal.v7i1.12562

Minor, K. A., & Kartowisastro, I. H. (2022). Automatic Music Transcription Using Fourier Transform for Monophonic and Polyphonic Audio File. Ingénierie Des Systèmes d Information, 27(4), 629–635. https://doi.org/10.18280/isi.270413

Nath, S., Shahi, A. K., Martin, T., Choudhury, N., & Mandal, R. (2024). A Comparative Study on Speech Emotion Recognition Using Machine Learning. https://doi.org/10.1007/978-981-99-5435-3_5

Wijaya, A. A., Yasmina, I., & Zahra, A. (2021). Indonesian Music Emotion Recognition Based on Audio with Deep Learning Approach. Advances in Science, Technology and Engineering Systems Journal, 6(2), 716–721. https://doi.org/10.25046/aj060283

Wunarso, N. B., & Soelistio, Y. E. (2017). Towards Indonesian speech-emotion automatic recognition (I-SpEAR). Proceedings of 2017 4th International Conference on New Media Studies, CONMEDIA 2017, 2018-January. https://doi.org/10.1109/CONMEDIA.2017.8266038

Zahra, H. N., Ibrohim, M. O., Fahmi, J., Adelia, R., Nur Febryanto, F. A., & Riandi, O. (2020). Speech emotion recognition on indonesian youtube web series using deep learning approach. 2020 5th International Conference on Informatics and Computing, ICIC 2020. https://doi.org/10.1109/ICIC50835.2020.9288650