Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

A BiLSTM-Based Approach For Speech Emotion Recognition In Conversational Indonesian Audio using SMOTE

Nariswari Nur Shabrina*[1], Fatan Kasyidi[2], Ridwan Ilyas[3]

[1,2,3]Computer Science, Universias Jenderal Achmad Yani, Indonesia

Email: [1]nariswari21@if.unjani.ac.id

Received : Jul 27, 2025; Revised : Aug 25, 2025; Accepted : Sep 5, 2025; Published : Oct 16, 2025

Abstract

Speech Emotion Recognition (SER) identifies human emotions through voice signal analysis, focusing on pitch, intonation, and tempo. This study determines the optimal sampling rate of 48,000 Hz, following the Nyquist-Shannon theorem, ensuring accurate signal reconstruction. Audio features are extracted using Mel-Frequency Cepstral Coefficients (MFCC) to capture frequency and rhythm changes in temporal signals. To address data imbalance, Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic data for the minority class, enabling more balanced model training. A One-vs-All (OvA) approach is applied in emotion classification, constructing separate models for each emotion to enhance detection. The model is trained using Bidirectional Long Short-Term Memory (BiLSTM), capturing contextual information from both directions, improving understanding of complex speech patterns. To optimize the model, Nadam (Nesterov-accelerated Adaptive Moment Estimation) is used to accelerate convergence and stabilize weight updates. Bagging (Bootstrap Aggregating) techniques are implemented to reduce overfitting and improve prediction accuracy. The results show that this combination of techniques achieves 78% accuracy in classifying voice emotions, contributing significantly to improving emotion detection systems, especially for under-resourced languages.

Keywords : *BiLSTM, Bootstrap Aggregating, Nadam, Nyquist Shannon , One-vs-All, SMOTE, Speech Emotion Recognition.*

1. INTRODUCTION

In daily life, humans do not solely rely on verbal communication but also employ vocal elements such as tone, pitch, and tempo to convey emotions and feelings. These elements significantly influence the understanding of a person's emotional state. Speech Emotion Recognition (SER) technology enables devices to detect human emotions through the analysis of voice signals, allowing systems to identify complex sound patterns, including variations in pitch, intonation, and sound quality affected by environmental noise. When computers are able to comprehend human emotions with a level of accuracy comparable to that of humans, this capability offers profound benefits for enhancing education, science, and technology. One notable advantage lies in the utilization of artificial intelligence to monitor mental health in an optimal manner, free from the constraints of accessibility [1]. This technology holds considerable promise for addressing the growing mental health issues among adolescents of compulsory education age [2].

In the context of SER, voice signals are analyzed to identify features such as pitch, intonation, and tempo, which are essential for detecting both the type and intensity of emotions [3]. Emotion, as an expression of human feelings, directly influences voice patterns, and its intensity plays a pivotal role in altering the voice. The efficacy of the SER system relies on its ability to extract relevant voice

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

features and the model's capacity to recognize emotions based on intricate voice patterns. These voice patterns exhibit temporal characteristics, involving changes in frequency and rhythm over time, which are intrinsically linked to the emotions being conveyed [5]. Consequently, in speech signal processing, the selection of an appropriate sampling rate is crucial to ensure the high quality of feature extraction. According to the Nyquist-Shannon theorem, to accurately reconstruct a voice signal, the sampling rate must be at least twice the highest frequency present in the original signal [6]. Once the appropriate sampling rate is determined, methods such as Mel-Frequency Cepstral Coefficients (MFCC) are employed to extract voice features and capture temporal changes, which are essential for identifying emotions by simulating the human auditory process [7]. The data is subsequently normalized using Min-Max Scaling to ensure that the extracted features have values ranging from 0 to 1 [8].

The identification of emotion classes occurs after the normalization process using machine learning techniques to process and recognize complex patterns in labeled data. Previous studies have demonstrated that to handle complex voice patterns, not only MFCC but also a combination of lexical and acoustic features are utilized. Recurrent Neural Network (RNN)-based models, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), have proven effective in capturing conversational context, with an accuracy of 72.52% for classifying four emotion classes [9]. These studies indicate that contextual feature processing using Indonesian conversational data is more effective in capturing emotional transitions compared to non-contextual approaches. LSTM, in particular, has demonstrated its ability to capture temporal patterns and contextual relationships across utterances, thereby facilitating better identification of emotional transitions [7]. Furthermore, the application of boosting techniques combined with Synthetic Minority Over-sampling Technique (SMOTE) to balance test data with 39 MFCC coefficients and a 22,050 Hz sampling rate resulted in an accuracy of 65%, which helps mitigate overfitting issues [10]. The Hybrid Sampling Method can also be implemented to enhance efficiency in the analysis of complex voice data. This method combines Bucher Experimental Design and Latin Hypercube Sampling (LHS) to generate samples that are more representative of a range of random variables and intervals. By integrating these two techniques, Hybrid Sampling allows for more efficient sampling, reducing the number of samples required without compromising the quality of analysis [11]. Additionally, the One-vs-All (OVA) method serves as a solution for handling multi-label classification, where each emotion class is treated as a separate binary classification, enabling the model to handle various emotions accurately and effectively, even with high data complexity and variance [12].

The use of Convolutional Neural Networks (CNN) has also proven effective in recognizing spatial patterns in both voice and facial expressions, with an accuracy of 88.71%, indicating the potential for multimodal data integration in emotion recognition [13]. However, due to the inherent limitations of CNNs in handling temporal patterns in sequential data such as voice, RNN-based models like LSTM and BiLSTM are more commonly employed. BiLSTM (Bidirectional Long Short-Term Memory) captures temporal context from both directions, allowing the model to better understand complex voice patterns and improve emotion identification [14]. This capability is further optimized when training processes utilize the Nadam algorithm (Nesterov-accelerated Adaptive Moment Estimation), which combines the benefits of the Adam optimizer with Nesterov's momentum. This optimizer accelerates convergence and enhances the stability of weight updates during the training process [15]. Furthermore, Esnemble Learning techniques, such as Bagging (Bootstrap Aggregating) are employed to enhance the model's performance and stability. Ensemble Learning combines predictions from multiple models trained on different subsets of the data to improve accuracy and reduce overfitting. Bagging is particularly effective in addressing data imbalance by allowing each model in the ensemble to focus on a different aspect of the data, thus reducing the overall variance and improving the generalization ability of the model. In the context of emotion recognition, Ensemble

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

Learning via Bagging helps to combine predictions from multiple models, ensuring that the final emotion classification benefits from the diversity of the models and is less prone to overfitting [15].

This study adopts an approach to emotion recognition in speech by integrating BiLSTM and MFCC for Indonesian conversational speech data. The use of Indonesian as the target language presents unique challenges, particularly due to the limited availability of labeled speech datasets. This lack of resources makes traditional methods for data balancing less effective, as the models often struggle to handle the imbalanced nature of the dataset. Several techniques for addressing data imbalance, such as oversampling and undersampling, may not be well-suited for the specific characteristics of Indonesian speech data. Therefore, this research aims to compare different data balancing methods to improve emotion class recognition performance. By analyzing how these methods enhance emotion detection in Indonesian speech, the study seeks to identify the most appropriate approach for achieving a well-balanced dataset that enhances model performance without compromising data integrity. Additionally, Ensemble Learning via Bagging is applied to address overfitting and improve the model's generalization ability.

2. METHOD

The experimental method is employed in this study to directly observe the effects of various stages involved in the voice signal analysis process, including data preprocessing, feature extraction using MFCC, data imbalance handling using Oversampling or Hybrid Sampling techniques, as depicted in Figure 1 at the data balancing stage, and modeling with BiLSTM and ensemble learning. The aim of this experimental research is to measure the effectiveness of the applied techniques in improving emotion classification accuracy based on voice signals.

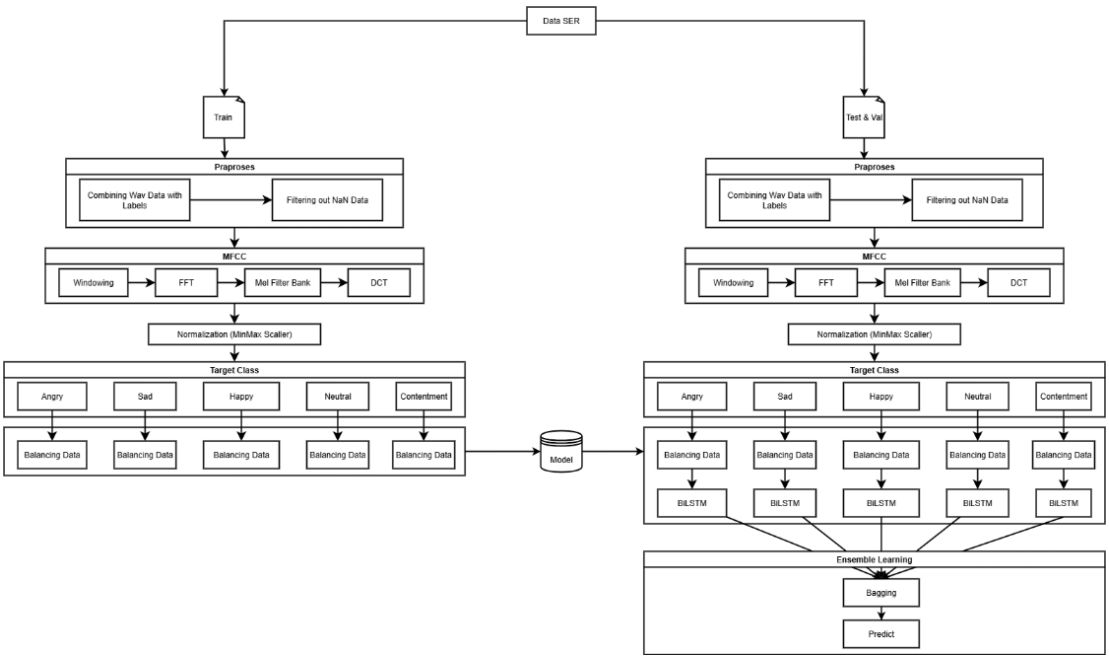

Figure 1. Method Research Flowchart

With an experimental approach, the performance of models using different methods can be compared and their impact on classification results analyzed. Observations are made by systematically testing the model through several stages, starting from feature extraction to the application of modeling techniques and ensemble learning. The results obtained from this experiment are then analyzed to determine whether the application of these techniques can improve emotion classification accuracy

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

and address the data imbalance issues that often occur in emotion-based voice datasets. Therefore, this approach allows the research to produce measurable and repeatable results, while identifying the factors that contribute to the improvement of model performance.

2.1. Dataset

The dataset used in this study was obtained from previous research and consists of recordings of Indonesian conversations collected from four television talk shows: Sarah Sechan, Indonesia Lawyers Club, Mata Najwa, and Tonight Show [10]. As shown in Table 2, the dataset consists of 3,784 samples; however, only 3,765 samples are usable, as shown in Table 1, due to some files being corrupted during the data collection process. However, due to the limited availability of annotated speech data in Indonesian, this dataset presents several challenges, including imbalanced emotion class, which are addressed using multiple techniques to handle the imbalance and mitigate overfitting.

Table 1. Distribution of Audio Files Across Different Talk Shows

Folder Name	Number of Files (.wav)
audio_tonightshow	787
audio_sarahsechan	1020
audio_matanajwa	1022
audio_ILC	936
Total Data Amount	**3765**

Table 2. Distribution of Emotion Classes in the Dataset

Emotion Class	Number of Data
Happy	391
Angry	234
Sad	287
Netral	2698
Contentment	174
Total Data Amount	**3784**

The audio signals exhibit varying lengths and are affected by noise disturbances caused by environmental factors. Each recording is labeled according to one of five emotion classes: Angry, Sad, Contentment, Neutral, and Happy. The distribution of emotion classes in this dataset is imbalanced, with the Neutral class containing a larger number of samples compared to the other classes. This imbalance presents a challenge during model training, making the handling of extreme data imbalance a critical aspect of this study. All collected audio recordings are paired with their corresponding labels to form a complete and structured dataset.

2.2. Feature Extraction Using MFCC

MFCC (Mel-Frequency Cepstral Coefficients) is a commonly used feature in speech signal processing due to its foundation on the Mel frequency scale, which aligns more closely with human perception of sound [7]. The MFCC extraction process begins by dividing the speech signal into small frames. Each frame is then windowed using a Hamming function to reduce the effects of discontinuity. Afterward, the FFT (Fast Fourier Transform) is applied to convert the signal into the frequency domain [17]. Next, a Mel filterbank is applied to adjust the frequency spectrum to align with human perception of sound, emphasizing relevant frequencies. The energy of these frequencies is then calculated and converted into coefficients that represent the features of the speech signal.To reduce the data dimensionality, the Discrete Cosine Transform (DCT) is applied, which produces the MFCC

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

coefficients used in model training. In this study, the number of MFCC coefficients used is 13, which is a common choice in speech signal analysis.

2.3. Data Normalization Using Min-Max Scalling

After feature extraction using MFCC, the data is then normalized using Min-Max Scaling. This process ensures that all feature values are within a uniform range, specifically between 0 and 1 [18]. This normalization is crucial because different features extracted from the speech signal may have varying scales, which can impact the model's performance during training. Without normalization, features with larger scales may dominate the model, while features with smaller scales may be overlooked. By using Min-Max Scaling, the feature values are forced to fall within the same range, allowing the model to learn from all features more fairly and effectively, while also improving the model's convergence during training. This process ensures that the model can optimize the weights for all features without bias towards those with larger values.

2.4. One-Vs-All for Multi-Class Classification

To handle multi-class classification, One-vs-All (OvA) is applied, where each emotion class is considered as the positive class (1), and all other classes are treated as the negative class (0). With this approach, the model is trained to predict whether a speech signal belongs to a specific emotion class or not. This process begins by building a binary classification model for each class. For instance, for the Happy class, the model will be trained to distinguish between speech signals that belong to the Happy category (positive, 1) and those that do not (negative, 0). The same process is applied to each of the other emotion classes, such as Sad, Angry, Contentment, and Neutral. Each model constructed will output either 0 or 1, indicating whether the speech signal belongs to the targeted emotion class or not. Once all models are trained, when a new speech signal is provided, each model will make a prediction. The class that receives an output of 1 from its model is considered the most likely emotion for the speech signal.This approach enables the model to effectively handle more than two classes, even though it only produces binary predictions for each class. The final output will be determined by the class that yields a prediction of 1, indicating the most probable emotion class for the analyzed speech signal. Thus, the One-vs-All technique allows for a clear distinction between different emotion classes, even within a single classification model [19].

2.5. Handling Data Imbalance with Oversampling

Before proceeding to the data balancing stage, the data is first split into two parts: 70% for training data and 30% for testing (test) or validation (val) data. After this division, data balancing is performed using SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic data for the minority classes in each emotion category present in the training data [20]. This process aims to address the imbalance in data between the majority and minority classes, allowing the model to be trained in a more balanced manner. This helps the model to more effectively identify all emotion classes, including those with fewer data points.

2.6. Handling Data Imbalance with Hybrid Sampling

In addition to using the oversampling technique with SMOTE, another experiment was conducted for data balancing using the hybrid sampling technique. In this approach, two different techniques are combined SMOTE oversampling on the minority class in the training data and Random Under-Sampling (RUS) undersampling on the majority class in the validation and test data [21]. Given that the dataset is relatively small and Hybrid Sampling is applied, the data is split 50% for training and 50% for validation. The SMOTE technique is used to increase the number of samples in the minority class of the training data, generating synthetic data that helps the model better recognize

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

emotional patterns. Meanwhile, RUS is applied to the majority class in the validation and test data to prevent the majority class from dominating by reducing data from the majority class, while keeping the validation and test data intact [22] [23].

2.7. Modelling Using BiLSTM

The BiLSTM model is used to process speech data and identify temporal patterns in the audio signals. BiLSTM was chosen for its ability to process data from both directions as shown in Figure 2, enabling the model to capture a better context in conversation.

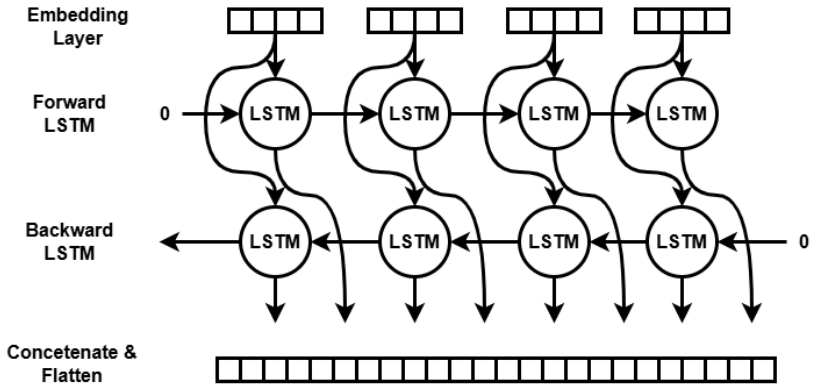

Figure 2. BiLSTM Model Architecture

In this experiment, the BiLSTM model is used for emotion classification based on speech signals. Each emotion class is treated as a separate model, where the model is trained to classify whether a speech signal belongs to a specific emotion class or not. This is done with two Bidirectional LSTM layers [24], each containing 128 units, to capture information from both directions in the data sequence. L2 regularization is applied to prevent overfitting, and Batch Normalization helps accelerate training. A dropout rate of 0.5 is used to reduce overfitting, and the output layer uses a sigmoid activation function for binary classification. Nadam with a learning rate of 1e-4 was selected as the optimizer for training stability. Two callbacks, EarlyStopping and ReduceLROnPlateau, are employed to stop training early if no improvement is observed and to reduce the learning rate if necessary. The model is trained with this setup to ensure stability and prevent overfitting. The results are then evaluated using a classification report and confusion matrix.

2.8. Ensemble Learning Bootstrap Aggregating

Bagging (Bootstrap Aggregating) is an ensemble learning technique that combines multiple models trained on different subsets of the data [25]. Each subset is created by random sampling with replacement. The individual model predictions are then aggregated as shown in Figure 3, typically by voting or averaging, to improve accuracy and reduce overfitting. In the context of the paper, Bagging enhances model stability and performance by leveraging the diversity of predictions from multiple models.

In this study, Bagging is used to combine the five models, each trained to classify a different emotion class, into a single unified model. By applying Bagging, the predictions from these five separate models are aggregated, which helps to reduce overfitting. This ensemble approach ensures that the final model benefits from the diversity of predictions across all emotion classes, improving both accuracy and generalization while minimizing the risk of overfitting that could arise from training individual models.

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

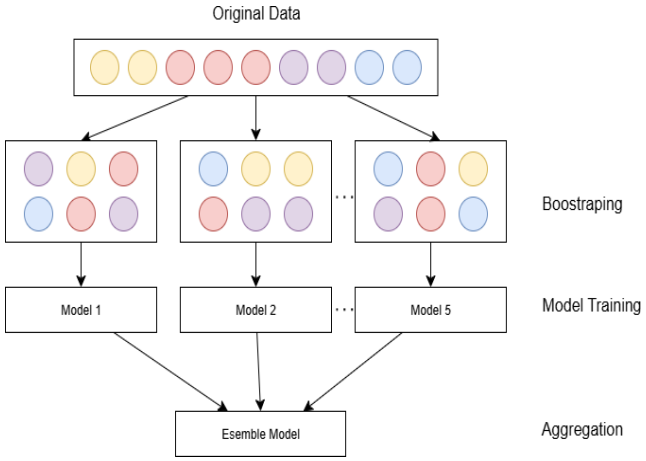

Figure 3. Bagging Architecture

3. RESULT

In this section, the experimental results conducted to evaluate the performance of the emotion recognition model based on voice signal analysis will be presented in detail. The experiment focuses on the application of various techniques to address data imbalance and overfitting in emotion classification. The techniques tested include One-Versus-All (OVA) for multi-class classification, SMOTE, Hybrid Sampling, and the Ensemble Learning approach using Bootstrap Aggregating.

3.1. One Vs All

This approach is used to classify multi-class problems by transforming them into binary classification tasks, where each emotion class represents a separate model. Each model is then trained to distinguish whether a given voice signal belongs to a specific emotion class or not. As shown in Table 3, one model will be trained to differentiate Neutral emotion (as the positive class) from other emotions (as the negative class), while another model will be trained to distinguish Sad from other emotions. This method applies to all emotion classes, such as Angry, Contentment, and Happy.

Table 3. One-Vs-All (OVA) Binary Classification Models for Emotion Recognition

Model	Emotion	Binary				
		Neutral	Sad	Angry	Contentment	Happy
1	Neutral	1	0	0	0	0
2	Sad	0	1	0	0	0
3	Angry	0	0	1	0	0
4	Contentment	0	0	0	1	0
5	Happy	0	0	0	0	1

Thus, each emotion has its own binary classification model. After these models are trained, test data is provided to each model, and the prediction results are obtained based on the model that outputs a positive result for the most relevant emotion class. OVA is one of the techniques used to address overfitting because the classification produced is clear and structured between emotion classes, although additional handling is required to manage data imbalance.

3.2. Oversampling

The oversampling stage in this experiment utilizes SMOTE (Synthetic Minority Over- sampling Technique). In the application of SMOTE, the data splitting technique employed is 70:30, where 70% of the data is used for training and 30% for testing and validation. This split is chosen to ensure that the

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

model receives sufficient data for training, while the validation and testing data remain unaffected. As shown in Table 4, 5, 6, 7, 8, this technique is applied exclusively to the training data, not to the validation or testing data, to ensure that the evaluation data still reflects the original class distribution without being influenced by oversampling. By adding synthetic samples to the minority class while keeping the validation and testing data intact, the evaluation process remains unbiased and based on the original data.

Table 4. SMOTE Technique Applied to Training Data (Neutral)

Split Data		Data	SMOTE
Train	Neutral (1)	1592	1592
	Non – Neutral (0)	714	1592
Test	Neutral (1)	165	165
	Non – Neutral (0)	329	329
Val	Neutral (1)	153	153
	Non – Neutral (0)	341	341

Table 5. SMOTE Technique Applied to Training Data (Sad)

Split Data		Data	SMOTE
Train	Sad (1)	109	109
	Non – Sad (0)	2197	109
Test	Sad (1)	30	30
	Non – Sad (0)	464	464
Val	Sad (1)	27	27
	Non – Sad (0)	467	467

Table 6. SMOTE Technique Applied to Training Data (Happy)

Split Data		Data	SMOTE
Train	Happy (1)	202	202
	Non – Happy (0)	2104	202
Test	Happy (1)	37	37
	Non – Happy (0)	457	457
Val	Happy (1)	40	40
	Non – Happy (0)	454	454

Table 7. SMOTE Technique Applied to Training Data (Angry)

Split Data		Data	SMOTE
Train	Angry (1)	242	242
	Non – Angry (0)	2064	242
Test	Angry (1)	62	62
	Non – Angry (0)	432	432
Val	Angry (1)	58	58
	Non – Angry (0)	436	436

Table 8. SMOTE Technique Applied to Training Data (Contentment)

Split Data		Data	SMOTE
Train	Contentment (1)	161	161
	Non – Contentment (0)	2145	161
Test	Contentment (1)	24	24
	Non – Contentment (0)	470	470
Val	Contentment (1)	40	40
	Non – Contentment (0)	454	454

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

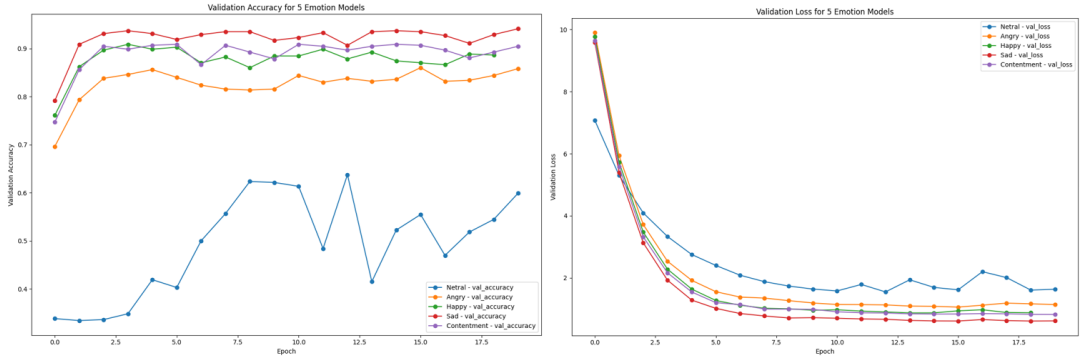

Figure 4. Model Performance Oversampling: Accuracy and Loss Comparison

This approach, as shown in Figure 4, positively impacts the stability of the model. Although the training data becomes more balanced, the unchanged validation and testing data help mitigate the risk of overfitting that could arise if the evaluation data were also altered. As a result, even though the number of training samples for the minority class increases, the model is still tested with more representative data, ensuring a more realistic evaluation

3.3. Hybrid Sampling

In the experiment using Hybrid Sampling (SMOTE + RUS), the data splitting technique applied is 50:50, where 50% of the data is used for training and 50% for testing and validation. This data split is chosen with the consideration that RUS reduces the number of majority class samples in the testing and validation sets while maintaining the integrity of the data. Therefore, by dividing the data into two balanced parts, the aim is to ensure that, despite the reduction of majority class data by RUS, the remaining data does not become insufficient, considering the relatively small overall dataset size. The Hybrid Sampling technique (SMOTE + RUS) is applied to the training, testing, and validation data as shown in Table 9, 10, 11, 12, 13. In this case, SMOTE is used to add synthetic samples to the minority class in the training data, while RUS is applied to reduce the number of majority class samples in the testing and validation data to maintain the authenticity of the data. As a result, the training, testing, and validation data are balanced.

Table 9. SMOTE + RUS Technique Applied to Training, Test, and Val Data (Neutral)

Split Data		Data	SMOTE + RUS
Train	Neutral (1)	1138	1138
	Non Neutral (0)	509	1138
Test	Neutral (1)	551	273
	Non Neutral (0)	273	273
Val	Neutral (1)	574	250
	Non Neutral (0)	250	250

Table 10. SMOTE + RUS Technique Applied to Training, Test, and Val Data (Sad)

Split Data		Data	SMOTE + RUS
Train	Sad (1)	80	1567
	Non Sad (0)	1567	1567
Test	Sad (1)	41	41
	Non Sad (0)	783	41
Val	Sad (1)	45	45
	Non Sad (0)	779	45

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

Table 11. SMOTE + RUS Technique Applied to Training, Test, and Val Data (Happy)

Split Data		Data	SMOTE + RUS
Train	Happy (1)	142	1505
	Non Happy (0)	1505	1505
Test	Happy (1)	62	62
	Non Happy (0)	762	62
Val	Happy (1)	75	75
	Non Happy (0)	749	75

Table 12. SMOTE + RUS Technique Applied to Training, Test, and Val Data (Angry)

Split Data		Data	SMOTE + RUS
Train	Angry (1)	1138	1138
	Non Angry (0)	509	1138
Test	Angry (1)	551	273
	Non Angry (0)	273	273
Val	Angry (1)	574	250
	Non Angry (0)	250	250

Table 13. SMOTE + RUS Technique Applied to Training, Test, and Val Data (Contentment)

Split Data		Data	SMOTE + RUS
Train	Contentment (1)	118	1592
	Non Contentment (0)	1592	1592
Test	Contentment (1)	50	50
	Non Contentment (0)	774	50
Val	Contentment (1)	57	57
	Non Contentment (0)	767	57

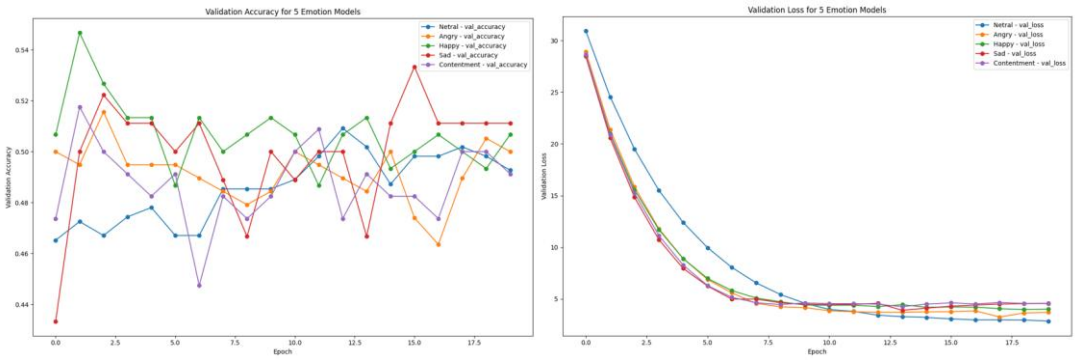

Figure 5. Model Performance Hybrid Sampling: Accuracy and Loss Comparison

Hybrid Sampling (SMOTE + RUS) aims to balance the dataset and prevent the dominance of the majority class during model training. However, the application of RUS to the validation and testing data results in significant fluctuations in accuracy as shown in Figure 5. While the distribution of the validation and testing data remains more natural, the reduction in the number of majority class samples leads to difficulties in generalization and underfitting, causing larger fluctuations in the performance of each emotion model.

3.4. Ensemble Learning

Ensemble Learning, in this context, specifically employs Bagging (Bootstrap Aggregating) to combine the strengths of multiple models, thereby enhancing both the accuracy and stability of the model in classifying emotions based on voice signal analysis. Bagging works by training several models on different subsets of the data, each obtained through random sampling with replacement. The

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

predictions from these individual models are then aggregated, typically through voting or averaging, to produce the final prediction. This process helps reduce the risk of overfitting by averaging out errors from individual models, thereby improving the model's generalization capability. As a result, Bagging enhances the reliability of emotion classification by leveraging the diversity of multiple models, enabling the system to handle the complexity and variability of the data more effectively.

3.4.1. Oversampling

In Ensemble Learning, using SMOTE as a data balancing technique, as shown in Figure 4, the model's performance with SMOTE demonstrates a stable improvement in both accuracy and loss compared to individual models. This approach helps address class imbalance by generating synthetic samples for the minority class, while the validation and testing data remain unaffected.

Table 14. Classification Report Ensemble Learning Bagging: SMOTE

	Precision	Recall	F1-Score
Neutral	0.7889	0.9208	0.8498
Sad	0.7692	0.3704	0.5000
Angry	0.7838	0.5000	0.6105
Contentment	0.6667	0.5500	0.6027
Happy	0.8000	0.5000	0.6154
accuracy			**0.7806**
macro avg	0.7617	0.5682	0.6357
weighted avg	0.7785	0.7806	0.7656

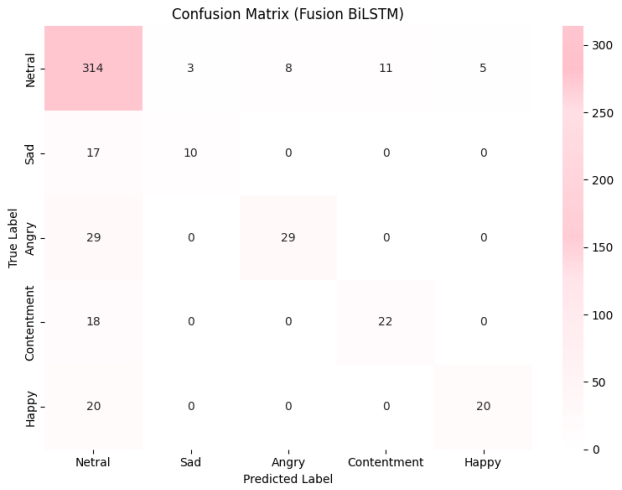

Figure 6. Confusion Matrix Ensemble Learning Bagging: SMOTE

Based on Table 14, the model demonstrates relatively good performance across several classes. The highest accuracy is recorded for the Neutral and Happy classes, with good precision and recall. However, the model's performance decreases for the Sad and Contentment classes, where both recall and precision are lower. For instance, in the Sad class, recall is very low at 0.3704, indicating that the model struggles to identify this emotion accurately. Similarly, for Contentment, while it has relatively high precision, the lower recall indicates challenges in consistently detecting this emotion. Based on Figure 6, although SMOTE helps improve performance for the Neutral class, some misclassifications still occur for the Sad and Angry classes, as reflected in the high number of false positives for these classes. Contentment also continues to experience frequent misclassifications, often being categorized as Neutral.

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

3.4.2. Hybrid Sampling

In Ensemble Learning, using SMOTE + RUS as a data balancing technique, significant fluctuations occur, as seen in Figure 5. This is due to the use of RUS on the validation and testing data, which reduces the number of majority class samples. As a result, the validation and testing data become increasingly limited, affecting the model's ability to generate consistent and stable predictions. While SMOTE successfully increases the number of minority class samples in the training data, the application of RUS to the validation and testing data leads to significant fluctuations in the model's performance, particularly in terms of accuracy and loss across several classes.

Table 15. Classification Report Ensemble Learning Bagging: SMOTE + RUS

	Precision	Recall	F1-Score
Neutral	0.9718	0.9640	0.9679
Sad	0.9459	0.8537	0.8974
Angry	0.8868	0.9691	0.9261
Contentment	0.8636	0.7600	0.8085
Happy	0.8615	0.9032	0.8819
accuracy			**0.9280**
macro avg	0.9059	0.8900	0.8964
weighted avg	0.9287	0.9280	0.9274

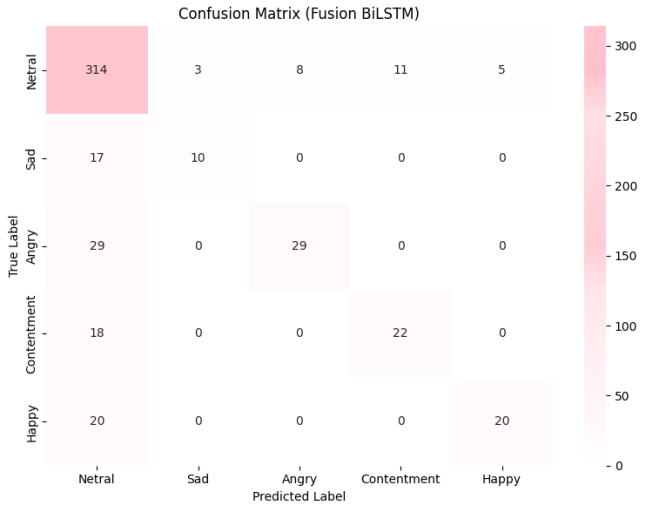

Table 7 Classification Report Ensemble Learning Bagging: SMOTE + RUS

Based on Table 15, the model demonstrates relatively good performance across all classes. For the Neutral class, the model achieves a very high precision of 0.9718 and recall of 0.9640, indicating the model's strong ability to accurately identify this emotion. The Sad class shows good precision 0.9459, but its recall is lower 0.8537. Based on Figure 7, although the overall accuracy reaches 0.9280, several emotion classes still experience misclassifications, particularly those classified as Neutral, which indicates that the model is still struggling to consistently differentiate between certain emotion classes.

4. DISCUSSIONS

The results of this study demonstrate that the application of the SMOTE technique significantly enhances the model's performance in classifying five emotion categories. In previous research, an accuract of 65% was achieved using SMOTE as data balancing technique [9]. By addressing the issue of data imbalance, SMOTE yields an accuracy of 0.7806, as illustrated by the generally favorable trend in the corresponding graph, which reflects stability in emotion classification. These findings are

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

consistent with those of [20], which suggest that SMOTE improves model performance by mitigating the dominance of the majority class, thereby enabling the model to more effectively recognize all emotion classes, especially those underrepresented in the dataset.

Furthermore, the combination of SMOTE and RUS results in an even higher accuracy of 0.9280, as reported in [26], which asserts that hybrid sampling is an effective strategy for tackling data imbalance. This approach combines SMOTE to augment the minority class samples and RUS to reduce the majority class samples. Despite the higher accuracy achieved, the combination of SMOTE and RUS yielded suboptimal performance in this study. The primary reason for this outcome is the relatively small size of the dataset, which renders the undersampling technique less effective, leading to significant fluctuations in model performance. The use of hybrid sampling also presents certain drawbacks, as the reduction of majority class data can compromise the model's stability. In this case, it resulted in underfitting for specific emotion classes, as highlighted by [11]. The RUS technique applied to the small dataset diminishes the variation in the data, which is essential for training a robust model, thus contributing to a reduction in both accuracy and consistency for some of the emotion classes.

5. CONCLUSION

The use of the SMOTE technique has proven to yield more stable and effective results. This technique allows for better model performance by maintaining data balance and ensuring that the model can learn effectively without compromising the integrity of the evaluation data. The model demonstrates consistent results with stable accuracy and loss, as shown in Figure 4. In contrast, the application of SMOTE + RUS while showing higher accuracy as seen in Table 7, experiences significant fluctuations, as shown in Figure 5. The use of RUS to reduce the number of majority class samples results in substantial fluctuations in recall and precision across several classes, particularly in the Sad and Contentment classes. The reduction in majority class data during testing impacts the model's ability to generate consistent and stable predictions, despite the overall higher accuracy. The model also demonstrates underfitting in certain classes, In contrast, the application of SMOTE + RUS, while showing higher accuracy as seen in Table 8, experiences significant fluctuations, as shown in Figure 5. The use of RUS to reduce the number of majority class samples results in substantial fluctuations in recall and precision across several classes, particularly in the Sad and Contentment classes. The reduction in majority class data during testing impacts the model's ability to generate consistent and stable predictions, despite the overall higher accuracy. The model also demonstrates underfitting in certain classes.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

The author would like to express gratitude to Universitas Jenderal Achmad Yani for support and the facilities provided, which greatly contributed to the success of this research. Thanks for the opportunity to conduct this study in such a conducive environment, and for all the assistance that made it possible to complete this research effectively.

REFERENCES

[1] Nelly Elsayed, Zag ElSayed, Navid Asadizanjani, Murat Ozer, Ahmed Abdelgawad, and Magdy Bayoumi, “Speech Emotion Recognition using Supervised Deep Recurrent System for Mental Health Monitoring,” Jun. 2023.

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

[2] Jegadeesan S, Aswin Kumar S, Madhan K, Karthick G, S Gowdhamkumar, and G Anushree, "Real Time Speech Emotion Recognition for Mental Health Monitoring," Apr. 2025.

[3] L. Yu, F. Xu, Y. Qu, and K. Zhou, "Speech emotion recognition based on multi-dimensional feature extraction and multi-scale feature fusion," *Applied Acoustics*, vol. 216, Jan. 2024, doi: 10.1016/j.apacoust.2023.109752.

[4] M. V. Subbarao, S. K. Terlapu, and P. S. R. Chowdary, "Emotion Recognition using BiLSTM Classifier," in *Proceedings - 2022 International Conference on Computing, Communication and Power Technology, IC3P 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 195–198. doi: 10.1109/IC3P52835.2022.00048.

[5] Z. Zeng, J. Liu, and Y. Yuan, "A Generalized Nyquist-Shannon Sampling Theorem Using the Koopman Operator," *IEEE Transactions on Signal Processing*, vol. 72, pp. 3595–3610, 2024, doi: 10.1109/TSP.2024.3436610.

[6] D. B. Riyanto, A. Y. Rahman, and Istiadi, "Children with Speech Disorders Voice Classification: LSTM and BiLSTM Approach Based on MFCC Features," in *Proceedings: ICMERALDA 2023 - International Conference on Modeling and E-Information Research, Artificial Learning and Digital Applications*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 35–38. doi: 10.1109/ICMERALDA60125.2023.10458192.

[7] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," Dec. 2022, doi: 10.1016/j.asoc.2022.109924.

[8] A. N. I. Adma and D. P. Lestari, "Conversational Speech Emotion Recognition From Indonesian Spoken Language Using Recurrent Neural Network-Based Model," in *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAICTA53211.2021.9640273.

[9] F. KASYIDI, R. ILYAS, and N. M. ANNISA, "Peningkatan Kemampuan Pengenalan Emosi Melalui Suara dalam Bahasa Indonesia," *MIND Journal*, vol. 6, no. 2, pp. 194–204, Dec. 2021, doi: 10.26760/mindjournal.v6i2.194-204.

[10] X. J. Meng, L. X. Zhang, Z. M. Liu, Y. Pan, and S. T. Zhu, "Hybrid sampling method for structural reliability analysis," in *Proceedings - 2020 International Conference on Artificial Intelligence and Computer Engineering, ICAICE 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 408–411. doi: 10.1109/ICAICE51518.2020.00086.

[11] E. Utami, Rini, A. F. Iskandar, and S. Raharjo, "Multi-Label Classification of Indonesian Hate Speech Detection Using One-vs-All Method," in *Proceedings - 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering: Applying Data Science and Artificial Intelligence Technologies for Global Challenges During Pandemic Era, ICITISEE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 78–82. doi: 10.1109/ICITISEE53823.2021.9655883.

[12] H. Avula, R. Ranjith, and A. S. Pillai, "CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions," in *6th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1360–1365. doi: 10.1109/ICECA55336.2022.10009316.

[13] M. Subramanian, S. Lakshmi Swetha, and V. R. Rajalakshmi, "Deep Learning Approaches for Melody Generation: An Evaluation Using LSTM, BILSTM and GRU Models," in *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCNT56998.2023.10308344.

[14] K. Chakrabarti and N. Chopra, "A State-Space Perspective on the Expedited Gradient Methods: Nadam, RAdam, and Rescaled Gradient Flow," in *2022 8th Indian Control Conference, ICC 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 31–36. doi: 10.1109/ICC56513.2022.10093397.

[15] Y. Shi, Z. C. Lin, J. Chen, X. Kang, Q. Yan, and C. Wei, "Research on Vibration Event Classification in Φ- OTDR Systems Using MFCC Feature Extraction and Improved Swin Transformer," in *2024 22nd International Conference on Optical Communications and Networks,*

Vol. 6, No. 5, October 2025, Page. 3173-3187
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5183

ICOCN 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICOCN63276.2024.10648329.

[16] S. Zhao, Y. Zhang, N. Xia, K. Zhang, J. Kuai, and Y. Zhang, "Research on Electricity Price Prediction Based on Combination Model," in *2024 3rd International Conference on Energy and Electrical Power Systems, ICEEPS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 753–758. doi: 10.1109/ICEEPS62542.2024.10693046.

[17] H. Kadi, T. Sourget, M. Kawczynski, S. Bendjama, B. Grollemund, and A. Bloch-Zupan, "Segmentation, and Numbering in Oral Rare Diseases: Focus on Data Augmentation and Inpainting Techniques," in *Proceedings - 2023 International Conference on Computational Science and Computational Intelligence, CSCI 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1358–1363. doi: 10.1109/CSCI62032.2023.00298.

[18] S. A. Rufus, N. A. Ahmad, Z. Abdul-Malek, and N. Abdullah, "Thunderstorm Prediction Model Using SMOTE Sampling and Machine Learning Approach," in *APL 2023 - 12th Asia-Pacific International Conference on Lightning*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/APL57308.2023.10182046.

[19] B. Xu, W. Wang, R. Yang, and Q. Han, "An Improved Unbalanced Data Classification Method Based on Hybrid Sampling Approach," in *2021 IEEE 4th International Conference on Big Data and Artificial Intelligence, BDAI 2021*, Institute of Electrical and Electronics Engineers Inc., Jul. 2021, pp. 125–129. doi: 10.1109/BDAI52447.2021.9515306.

[20] A. S. Palli, J. Jaafar, M. A. Hashmani, H. M. Gomes, and A. R. Gilal, "A Hybrid Sampling Approach for Imbalanced Binary and Multi-Class Data Using Clustering Analysis," *IEEE Access*, vol. 10, pp. 118639–118653, 2022, doi: 10.1109/ACCESS.2022.3218463.

[21] T. Miyata, D. Kanemoto, and T. Hirose, "Random Undersampling Wireless EEG Measurement Device using a Small TEG," in *Proceedings - IEEE International Symposium on Circuits and Systems*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ISCAS46773.2023.10181822.

[22] H. Cui, L. Zhang, W. Wu, and Y. Peng, "A two-layer BiLSTM model with linear gating for Chinese named entity recognition," in *Proceedings of the International Joint Conference on Neural Networks*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/IJCNN54540.2023.10191631.

[23] M. Doostparast, M. Pouyani, and M. H. Y. Moghaddam, "Bootstrap Aggregating as an ensembled machine learning algorithm for power consumption prediction under asymmetric loss with linear model-base learners," in *2023 27th International Electrical Power Distribution Conference, EPDC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 6–11. doi: 10.1109/EPDC59105.2023.10218875.

[24] H. Hairani, T. Widiyaningtyas, D. D. Prasetya, I. Saifudin, and A. Tholib, "Reducing Class Imbalance with Undersampling for Improvement of Classification Method in Liver Disease Classification," in *2024 Beyond Technology Summit on Informatics International Conference, BTS-I2C 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 171–175. doi: 10.1109/BTS-I2C63534.2024.10941826.

[25] M. M. Santoni, T. Basaruddin, K. Junus, and O. Lawanto, "Automatic Detection of Students' Engagement During Online Learning: A Bagging Ensemble Deep Learning Approach," *IEEE Access*, vol. 12, pp. 96063–96073, 2024, doi: 10.1109/ACCESS.2024.3425820.