Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

Integration of BERT-VAD, MFCC-Delta, and VGG16 in Transformer-Based Fusion Architecture for Multimodal Emotion Classification

Fisan Syafa Nayoma*[1], Kusnawi[2]

[1,2]Informatics, Universitas Amikom Yogyakarta, Indonesia

Email: [1]fisansyafa812@students.amikom.ac.id

Received : Jun 19, 2025; Revised : Jul 22, 2025; Accepted : Jul 31, 2025; Published : Aug 24, 2025

Abstract

Emotion is a condition that plays an important role in human interaction and is the main focus of intelligence research in utilizing multimodal. Previous studies have classified multimodal emotions but are still less than optimal because they do not consider the complexity of human emotions as a whole. Although using multimodal data, the selection of feature extraction and the merging process are still less relevant to improving accuracy. This study attempts to categorize emotions and improve precision through a multimodal methodology that utilizes Transformer-based Fusion. The data used consists of a synthesis of three modalities: text (extracted through BERT and assessed through the affective dimensions of NRC Valence, Arousal, and Dominance), audio (extracted through MFCC and delta-delta^2 from the RAVDESS and TESS datasets), and images (extracted through VGG16 on the FER-2013 dataset). The model is built by mapping each feature into an identical dimensional representation and processed through a Transformer block to simulate the interaction between modalities, known as feature-level interactions. The classification procedure is run through a dense layer with softmax activation. Model evaluation was performed using Stratified K-Fold Cross Validation with k=10. The evaluation results showed that the model achieved 95% accuracy in the ninth fold. This result shows a significant improvement from previous research at the feature level (73.55%), and underlines the effectiveness of the combination of feature extraction and Transformer-based Fusion. This study contributes to the field of emotion-aware systems in informatics, facilitating more adaptive, empathetic, and intelligent interactions between humans and computers in practical applications.

Keywords: *BERT, MFCC, Multimodal Emotion, NRC-VAD, Transformer-Based Fusion, VGG16*

1. INTRODUCTION

The World Health Organization (WHO) reported a rise in mental disorders, including anxiety and depression, in recent years. The WHO reported a 25% increase in the prevalence of anxiety and depression disorders in 2020. This rise is mostly linked to the global effects of the COVID-19 epidemic. This event illustrates the significance of comprehending the emotional essence of humanity. Consequently, there exists a necessity for technology that facilitates human emotional identification [1].

The system's capacity to identify and react to patient emotions is becoming increasingly significant in the realm of human-computer interaction (HCI). The recognition of emotions enables the system to modify its responses based on the patient's emotional state. Nevertheless, numerous impediments hinder the precision of emotion recognition. Occasionally, patients exhibit confused feelings. Recognizing emotions solely through text or voice frequently fails to capture the intricacies of complicated human emotions [2], [3].

A multimodal technique integrates text, audio, and visual data to discern emotions [4], [5]. Prior research has demonstrated that the incorporation of many modalities from different sources enhances system accuracy [6]. The study MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering by Khan et al. [7] presents a model that integrates BERT text and visual

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

representations from CNN independently before merging them using a transformer or called merging at the feature-level. The findings of this study demonstrated a notable enhancement in the accuracy of utilizing modality data sourced from various origins. The accuracy results achieved were 73.55% with 3 datasets (trimodal) namely Question, Visual, and Subtitles, significantly surpassing previous studies that reported figures between 44.42% and 51.83% with only 2 datasets (bimodal) namely Question and Visual. This indicates that integrating three modalities from various sources can enhance comprehension and improve applications. Consequently, the application of multimodal approaches is widely sought after in contemporary research [8], [9], [10].

The Bidirectional Encoder Representations from Transformers (BERT) model is commonly employed in text data due to its proficient comprehension of semantic context [11]. BERT specializes in understanding word meanings through the interactions of words inside phrases or across sentences, making it particularly effective at collecting complex emotional nuances in text data [12], [13]. Affective representations, such as Valence–Arousal–Dominance (VAD), are crucial for capturing nuanced emotional characteristics. VAD categorizes emotions into three psychological dimensions: valence, arousal, and dominance [14]. This study used a combination of BERT and VAD extraction based on these concerns.

The Mel-Frequency Cepstral Coefficients (MFCC), Delta-MFCC, and Delta^2-MFCC models in audio data encapsulate the acoustic properties of speech signals [15]. MFCC accurately captures spectral patterns linked to human vocal articulation. Delta-MFCC and Delta^2-MFCC captures temporal variations, enabling the detection of emotional dynamics such as intonation and vocal intensity [16], [17]. Voice-based emotion identification systems have extensively utilized this feature extraction method [18]. This study employs the MFCC and delta-MFCC models based on these concerns.

The VGG16 architecture serves as the visual feature extractor for facial expressions in picture data. VGG16 possesses the capability to discern spatial patterns, including eyebrow movements, smiles, and tense emotions [19], [20]. These visual characteristics are crucial as facial expressions serve as the primary markers of emotion in non-verbal communication. A study by Saravanan et al. [21] shown the successful application of VGG16 for face mask detection. This study used the VGG16 model based on these assumptions.

Evaluation necessitates the execution of multimodal data fusion. Conventional fusion methods, such as early and late fusion, frequently fail to capture intricate relationships between modalities. Transformer-Based Fusion was developed to address these limitations [22], [23]. This method employs self-attention and cross-attention mechanisms, along with Stratified K-Fold Cross Validation, to dynamically integrate diverse features [24], [25], [26]. This approach facilitates a more comprehensive understanding of the patient's psychological background.

Therefore, this study develops and evaluates an emotion classification model employing a Transformer-Based Fusion architecture. By effectively integrating multimodal data, the primary objective is to achieve a significant improvement in predictive accuracy compared to prior research. The findings are anticipated to contribute to the technological advancement of more nuanced and accurate emotion recognition systems.

2. METHOD

This research comprises four principal stages as illustrated in Figure 1, namely data collection (both primary and secondary), data preprocessing (encompassing text preprocessing, feature extraction, data normalization, and feature selection), development of a test system via multimodal data fusion utilizing Transformer-Based Fusion and the Multi-Head Self-Attention model, and evaluation to assess model performance.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

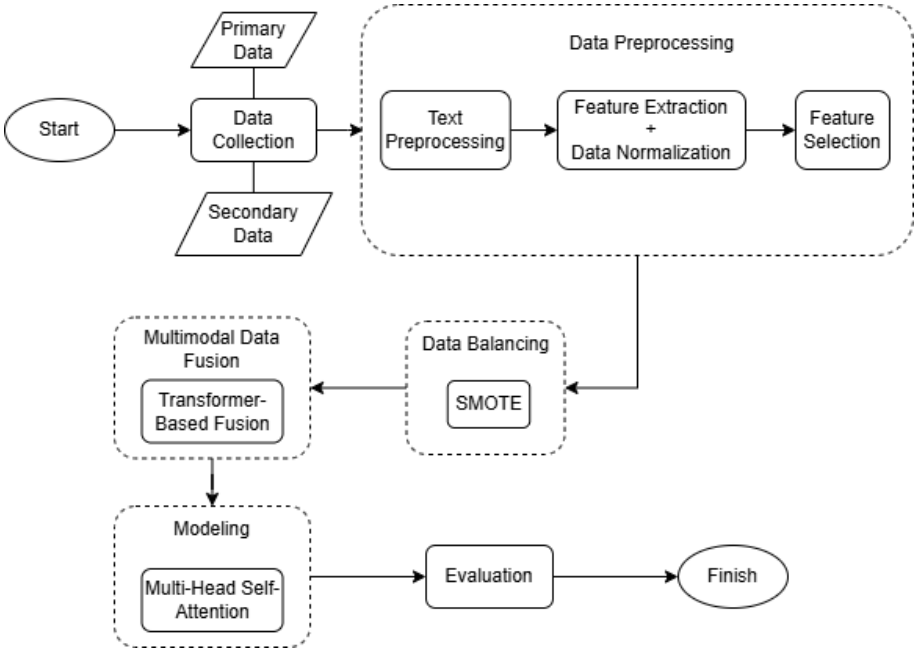

Figure 1. Research Stages

2.1. Data Collection

The main data used for multimodal emotion classification uses text, audio, and image modalities taken from Kaggle separately from tweet emotions, RAVDESS, TESS, dan FER-2013 data. FER-2013 was chosen in this study because until 2020 and beyond, this dataset is still often used to conduct facial image research [27], [28], [29]. Then, the supporting data used is VAD which comes from the NRC VAD Lexicon with updated version 2022. The supporting data is used to extract text features in order to capture affective dimensions [30].

2.2. Data Preprocessing

The primary stage in this study begins with selecting the target class to preserve data alignment across modalities. Given the multimodal nature of the data, it is crucial to verify that each modality has a consistent representation of the emotion class [31]. This technique is carried out by finding the target class available in each modality, then deleting any entries that do not have the same class in one of the modalities. In addition, to preserve semantic consistency, numerous emotion labels that have comparable meanings but are written differently will be standardized into one label [32].

2.2.1. Text Preprocessing

In the text modality, a preprocessing technique is carried out to remove noise and simplify the data structure. This stage includes converting all text to lowercase, eliminating digits, deleting punctuation, removing stopwords, and the stemming process. This approach seeks to provide cleaner and more representative data, hence boosting the effectiveness of the feature extraction process and subsequent analysis [33].

2.2.2. Feature Extraction

Feature extraction is done out independently according to the type of modality. In the text modality, a combination of BERT and NRC-VAD is utilized to capture phrase context and affective dimensions of emotion based on valence, arousal, and dominance [7]. The audio modality is retrieved using Mel Frequency Cepstral Coefficients (MFCC), Delta-MFCC, and Delta^2-MFCC. MFCC is employed to describe the spectral features of the speech signal, while Delta and Delta2-MFCC record

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

temporal variations to reflect the emotional dynamics of voice intonation [34]. The visual modality (picture) is retrieved using a pretrained VGG16 model, which is able to distinguish spatial patterns in facial emotions such as eyebrow movements, grins, and facial muscular tension [29].

All extracted features are then standardized using StandardScaler to guarantee that the numerical distribution between features is on a comparable scale [35]. This stage is necessary to improve the performance of the classification model and prevent the dominance of one characteristic over the others.

2.2.3. Feature Selection

Furthermore, a feature selection stage is carried out to increase the efficiency and accuracy of the model by filtering the features that are most relevant to the target label. This study uses the Mutual Information approach, which calculates the degree of reliance between each feature and the target class. The features with the highest mutual information values were picked because they were believed to have the most significant impact to the emotion classification process [36]. Unlike linear correlation methods such as Pearson, Mutual Information can detect complex relationships hidden inside multimodal data. This strategy improves the relevance of features to the target class while significantly reducing input dimensionality without sacrificing critical information [37].

2.3. Data Balancing

This study employed SMOTE for data balancing. The Synthetic Minority Over-sampling Technique (SMOTE) is an approach employed to address data imbalance by augmenting the minority class through the generation of synthetic data derived from existing minority data. In SMOTE, the selection of minority class instances and the determination of k-nearest neighbors for each instance are involved in the oversampling procedure. Subsequently, rather than merely replicating the current minority class instances, synthetic instances are generated through the combination or interpolation of the chosen instances. Thus, SMOTE can assist in addressing the issue of significant overfitting [38].

2.4. Modeling

This study models multimodal data using the Transformer-Based Fusion architecture. The architecture aims to balance the feature representation of the three modalities by utilizing the attention mechanism to identify interactions between modalities [39]. Each modality is processed separately through a 128-dimensional Dense layer with the ReLU activation function to equalize the dimensions of the feature space with the equation 1:

$$ModalityProj_i = ReLU(X_i W_i + b_i)\ For\ i\ \in \{text, audio, image\} \quad (1)$$

The ModalityProj or modality representations of text, audio, and image are combined into a fixed three-dimensional tensor to form an input structure of (batch_size, 3, 128) where X is input feature matrix, W is weight matrix, and b is bias vector. The Multi-Head Attention layer, which has four heads (num_heads = 4) and a key dimension of 64, collects cross-modality relations by the equation 2:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (2)$$

With Q = XWQ, K = XWK, V = XWV, and dk = 64 where the Q is query matrix, K is key matrix, and dk is dimesionality of key vectors. The results of several attention heads are combined through concatenation and linear projection with the equation 3:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \quad (3)$$

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

The output of this layer is recombined through residual connections and normalized using Layer Normalization, according to the basic principles of the Transformer encoder architecture with the equation 4:

$$Z_1 = LayerNorm\big(X + MultiHead(Q, K, V)\big) \quad (4)$$

Next, the attention results are processed by a two-layer feedforward neural network (FFN) block, each consisting of 128 neurons with ReLU activation using the equation 5:

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2 \quad (5)$$

To keep the activation distribution stable and facilitate better training, this process also uses residual connection and Layer Normalization after FFN using the equation 6:

$$Z_2 = LayerNorm\big(X + FFN(Z_1)\big) \quad (6)$$

To generate probabilities for each emotion class, Z_2 where Z is fused representation then summarized using Global Average Pooling with the equation 7:

$$z = \frac{1}{3}\sum_{i=1}^{3} Z_2[i] \quad (7)$$

This results in a single latent representation that represents the combined information from all three modalities. This representation is then mapped to the output space using a softmax-enabled Dense layer using the following equation 8:

$$\hat{y} = \text{softmax}(zW + b) \quad (8)$$

So, Transformer-Based Fusion not only combines modalities, but also models complex relation learning [40].

2.5. Evaluation

2.5.1. Validation Strategy

This study utilized a Stratified K-Fold Cross Validation strategy with 10 folds to ensure the validity and reliability of the performance results. This method partitions the complete dataset into 10 distinct, equally sized subsets, commonly referred to as "folds." This validation method employs 10 folds, allocating 90% of the data for training and 10% for testing. The validation process is conducted ten times, with each iteration holding one unique fold as the test set, while the other nine folds are combined to create the training set. As a result, every data point is utilized as test data one time and as training data nine times throughout all iterations.

An essential element of this approach is stratification, which rigorously preserves the same class distribution within each fold as it exists in the overall dataset. Stratification guarantees that every fold accurately reflects the class proportions of the dataset, effectively reducing the sampling bias that may occur with basic random splitting. This enhances the dependability of the overall performance assessments and offers a more precise evaluation of the model's capacity to generalize to novel, unseen data [41].

2.5.2. Evaluation Matrix

At this stage, perform an evaluation matrix by assessing the model performance through the use of a confusion matrix, as illustrated in Table. 1:

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

Table 1. Confusion Matrix

Classification	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

In using confusion matrix, there are 4 components that must be considered, namely true positive (TP) or positive data that is predicted to be correct, true negative (TN) or negative data that is predicted to be correct, false positive (FP) or negative data but predicted as positive, and false negative (FN) or positive data but predicted as negative. The data is also used to perform classification reports by calculating the accuracy, precision, recall, and f1-score values which can be seen in equations 9, 10, 11, and 12:

- Accuracy is the ratio of correct predictions from all data.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (9)$$

- Precision is the ratio of correct positive predictions from all predictions that result in positive.

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

- Recall is the ratio of correct positives from all data that are true positive.

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

- F1-Score is a comparison of precision and recall.

$$F1 - Score = \frac{Precision \times Recall}{Precision+Recall} \quad (12)$$

So, with the formula above, the results of the evaluation will be informative and can decide the next stage by looking at the classification report and confusion matrix [42].

3. RESULT

A sequence of procedures has been conducted to assess the efficacy of the multimodal emotion categorization model utilizing three modalities: text, audio, and visuals. This procedure entails assessing classification accuracy, analyzing the distribution of classifications among emotion labels, and evaluating the performance of each modality in relation to the final outcome following the fusion process utilizing the Transformer-Based Fusion architecture.

3.1. Data Collection

The utilized data comprises primary data and supporting data. The primary data comprises a dataset of three modalities: text with 44.001 data, audio with 4.240 data, and image with 35.900 data. The multimodal data has six selected categories: neutral, happy, sad, surprised, angry, and fear. The audio and image dataset is contained within a ZIP folder, with file names according to their respective sentiments or labels. The dataset for text data is structured as a CSV file, comprising columns for text and sentiment, which include a compilation of tweets that can be seen in the Table 2:

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

Table 2. Text Dataset

Feature	Description
Text	A collection of text taken from tweets that represent emotions.
Sentiment	Shows whether tweets in the text feature fall into one of the labels (neutral, happy, sad, surprised, angry, or fear).

For supporting data, utilize the VAD list of English words sourced from an emotion lexicon provider with updated version 2022.

3.2. Data Preprocessing

At this stage, each label is ensured to have the same name across modalities by performing label mapping. The goal is to ensure labels are consistent and there are no duplicate labels with the same intent but different label names. Label mapping provisions can be seen in Table 3:

Table 3. Label Mapping

Label name list	Label name standard
anger, angry	angry
fear, fearful	fear
happy, happiness	happy
neutral	neutral
sad, sadness	sad
surprise, surprised	surprised

Subsequently, it is essential to perform label equalization to guarantee that each modality has a uniform label, with the selected labels stored in the COMMON_LABELS variable: “neutral”, “happy”, “sad”, “surprised”, “angry”, and “fear”.

3.2.1. Text Preprocessing

At this stage, preprocess the text by converting to lowercase, removing numbers, removing punctuation, removing stopwords, and stemming. So, the results are as in Table 4:

Table 4. Text Preprocessing

No.	Before text preprocessing	After text preprocessing
1	the poetry event was a success. I don't think ...	poetri event success dont think ever realli th...
2	@MichelleZen That sounds good, too!	michellezen sound good
3	its sad that the rats are becoming aggressive ...	sad rat becom aggress guinea pig seper
4	I FOUND A PROM DRESS	found prom dress
5	got THE best mothers day present from Tys. It ...	got best mother day present ty made cri uncont...

With text preprocessing, the data will be cleaner, simpler, and standardized to create raw text that is easier and more effective to analyze.

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

3.2.2. Feature Extraction

Subsequently, feature extraction is conducted for each modality. In the text modality, employing the VAD Lexicon to quantify valence, arousal, and dominance on a scale from 0 to 1. Furthermore, the BERT model serves as a tool for text feature extraction. After that, perform data normalization using StandardScaler on the feature extraction results. Thus, it produces 771 feature columns in the text as in the Table 5:

Table 5. Text Feature Extraction

No.	Valence	Arousal	Dominance	BERT_0	BERT_1	…	BERT_767
1	0.664875	0.437500	0.613555	-1.519009	-0.150525	…	-1.498952
2	0.000000	0.000000	0.000000	0.137627	0.811491	…	-0.997816
3	0.407000	0.554359	0.459362	0.564261	0.407000	…	0.347465
4	0.000000	0.000000	0.000000	-0.033311	-2.065822	…	-1.617836
5	0.575000	0.385504	0.491770	-2.705802	-0.175697	…	-0.035887

In audio modality, employing MFCC, Delta-MFCC, and Delta^2-MFCC models. MFCC is selected to extract pertinent frequency data from the sound spectrum. Subsequently, Delta-MFCC and Delta^2-MFCC are employed to track variations in MFCC features across time intervals. After that, perform data normalization using StandardScaler on the feature extraction results. Thus, it produces 40 feature columns in the audio as in the Table 6:

Table 6. Audio Feature Extraction

No.	MFCC_0	MFCC_1	MFCC_2	MFCC_3	…	Delta2_MFCC_12
1	-1.193781	0.608742	0.227484	-0.360233	…	-0.134319
2	-1.201477	0.664056	0.701380	0.735217	…	-0.257554
3	0.043896	0.120376	-2.994666	-1.201312	…	-0.157851
4	-0.910055	-0.571925	-0.854156	-1.041108	…	0.010721
5	-0.889016	-0.612343	-1.641003	-1.194173	…	-0.348083

In the image modality, using the VGG16 model. This model is chosen to extract images into numeric feature vectors by capturing edges, textures, patterns, and shapes of objects. After that, perform data normalization using StandardScaler on the feature extraction results. Thus, it produces 512 feature columns in the image as in the Table 7:

Table 7. Image Feature Extraction

No.	0	1	2	3	…	510	511
1	-0.761640	-0.10832	0.064264	0.068542	…	0.005430	-0.308691
2	1.325315	-0.10832	-0.517475	3.167578	…	-0.172866	-0.308691
3	0.813441	-0.10832	2.568999	-0.371058	…	0.730289	-0.308691
4	0.872871	-0.10832	-0.231355	-0.908448	…	0.115922	-0.308691
5	-0.761640	-0.10832	0.064264	0.068542	…	0.005430	-0.308691

3.2.3. Feature Selection

from the feature extraction results, features that have a large number of columns will be selected using Mutual Information. In the text feature, there are 771 columns. Because there are too many features, feature selection is carried out on text by determining the 100 best features, the results of which can be seen in the Table 8:

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

Table 8. Text Feature Selection

No.	Valence	Arousal	Dominance	BERT_0	BERT_2	…	BERT_760
1	0.664875	0.437500	0.613555	-1.519009	-1.229215	…	1.335034
2	0.000000	0.000000	0.000000	0.137627	-0.204385	…	-1.419773
3	0.407000	0.554359	0.459362	0.564261	-1.913765	…	-1.577483
4	0.000000	0.000000	0.000000	-0.033311	-1.864708	…	1.014428
5	0.575000	0.385504	0.491770	-2.705802	-0.410916	…	0.844633

In addition, the image feature also produces a very large number of features reaching 512 columns. Therefore, the image feature also needs to be selected by selecting the 100 best features from all image features. The results can be seen in the Table 9:

Table 9. Image Feature Selection

No.	0	8	15	27	…	499	506
1	-0.761640	-1.196760	0.644437	-0.421204	…	-1.043551	0.324348
2	1.325315	-0.690800	-0.519002	-0.795902	…	-0.983943	-0.725453
3	0.813441	0.645709	0.107909	-0.400142	…	0.655538	2.126331
4	0.872871	0.172662	-0.601962	0.391708	…	0.120511	-1.305351
5	-1.829473	-1.275732	0.723356	-0.299732	…	-1.051156	-1.181131

For audio, it has 40 feature columns and is still relatively small so there is no need to select the features. Upon finalizing the feature selection process, the subsequent step involves executing modality alignment. The data quantity is established for each modality by identifying the minimum data set across all modalities, resulting in a uniform amount of 2856 per modality.

3.3. Data Balancing

This study utilized SMOTE to rectify data imbalance. The Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling method employed to correct class distribution by creating synthetic instances for minority classes derived from existing minority data. Before implementing SMOTE, the sample count per label for each modality was standardized by choosing the least sample size available among the three modalities. This approach ensures that each modality provides data uniformly, preventing any deficiency and preserving consistency among all input sources. The pre-balancing sample counts per label were 592 for the angry and happy classes, 496 for neutral, and 392 for sad, fear, and surprised. SMOTE was subsequently employed on each modality separately to produce synthetic data for underrepresented classes, yielding a balanced dataset. The ultimate data distribution following SMOTE processing is depicted in Figure 2:

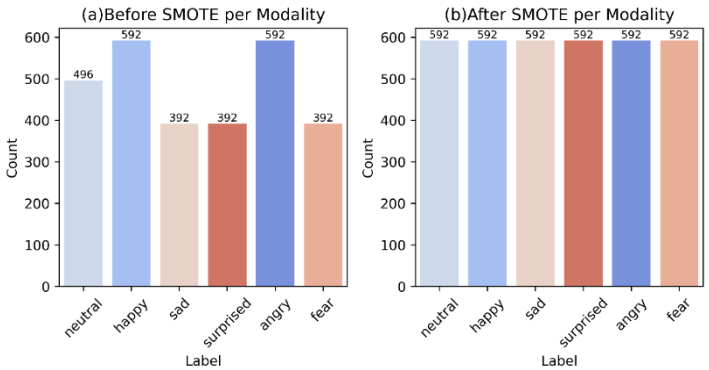

Figure 2. (a) before SMOTE per Modality; (b) after SMOTE per Modality

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

In Figure 2 indicates that each modality, following SMOTE application, is balanced with a total of 3552 data points per modality.

3.4. Modeling

This step builds a Transformer-based Fusion categorization architecture to combine text, audio, and image information with a total of 3552 data per modality after the SMOTE process. A feature vector extracted using BERT + NRC-VAD for text, MFCC, Delta-MFCC, and Delta^2-MFCC for audio, and VGG16 for image represents each modality. In Keras, the Input layer receives three inputs, one for each modality's feature dimension. Each feature vector is projected to the same spatial dimension using a 128-dimensional Dense layer and ReLU activation before fusion. The attention layer processes a 3D tensor created by combining the three projection representations using tf.stack.

Inter-modality fusion uses Multi-Head Self Attention with 4 heads (num_heads = 4) and 64 key dimensions. In an integrated spatial representation, this attention mechanism lets the model record contextual links between information from the three modalities. The residual connection from attention is added to the initial input and normalized using LayerNormalization. Next, output is sent to a two-layer Feed Forward Network (Dense 128 to ReLU, back Dense 128, then to ReLU again) with residuals and renormalization. After spatial aggregation with GlobalAveragePooling1D, a Dense layer with softmax activation classifies emotions.

3.5. Evaluation

3.5.1. Validation Strategy

The evaluation of model performance was performed using the Stratified K-Fold Cross Validation method with 10 folds. This strategy was selected to maintain a balanced distribution of class labels in each fold, hence rendering the model performance evaluation more representative.

The model is trained on each fold with predetermined parameters: batch size of 32, 15 epochs, and the Adam optimizer. Following the training on each fold, predictions are generated for the test data, after which the assessment metrics, specifically accuracy and weighted F1-score, are computed. Furthermore, a classification report and confusion matrix are generated for each fold to assess the model's efficacy across each emotion category with the report classification results in Table 10:

Table 10. Classification Report of Trimodal Best Fold

Fold	param	Precision	Recall	F1-Score	Support	Accuracy
9	Neutral	0.98	0.98	0.98	59	0.95
	Happy	0.98	0.97	0.97	60	
	Sad	0.95	0.98	0.97	59	
	Surprised	0.92	0.92	0.92	59	
	Angry	0.93	0.90	0.91	59	
	Fear	0.92	0.93	0.92	59	

The trimodal model's performance underwent a thorough evaluation through a 10-fold cross-validation approach. Table 10 provides a comprehensive overview of the classification metrics for the single fold that achieved the best performance, designated as Fold 9. The fold reached a peak accuracy of 0.95, which corresponds to 95%. The performance across the other folds exhibited variability, with the following accuracies recorded: 85% for fold 1, 81% for fold 2, 81% for fold 3, 88% for fold 4, 90% for fold 5, 91% for fold 6, 93% for fold 7, 94% for fold 8, and 93% for fold 10. Fold 9 is the best fold because it is the standard result of the cross-validation process that representing the model's peak capability on the specific data partition that proved most useful during iterative validation.

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

3.5.2. Evaluation Matrix

From the classification report from Table 10, confusion matrix of best fold trimodal model can be calculated which can be seen in Figure 3:

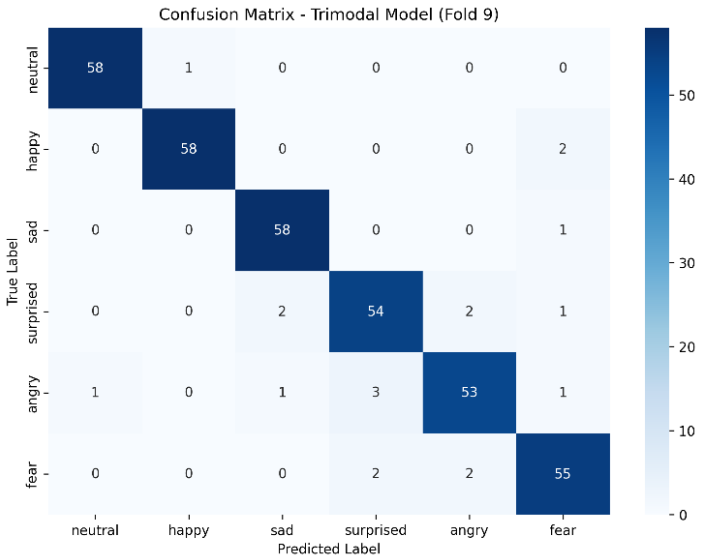

Figure 3. Confusion Matrix – Best Fold Trimodal Model

From the Figure 3, it is known that the model successfully classified 58 data as neutral correctly, 58 data as happy correctly, 58 data as sad correctly, 54 data as surprised correctly, 53 data as angry correctly, and 55 data as fear correctly.

Table 11. Comparison Between Models - Highest Accuracy

Models	Accuracy	F1-Score
BERT-VAD + MFCC, Delta, and Delta^2 + VGG16 + Early Fusion	0.93	0.93
BERT-VAD + MFCC, Delta, and Delta^2 + VGG16 + Late Fusion	0.94	0.94
BERT-VAD + MFCC, Delta, and Delta^2 + VGG16 + Transformer-Based Fusion	0.95	0.95

From Table 11, a comparative analysis was conducted to assess the efficacy of the proposed architecture, focusing on three distinct multimodal fusion strategies: Early Fusion, Late Fusion, and the proposed Transformer-Based Fusion. Table 11 illustrates that this experiment was structured to examine the influence of the fusion mechanism on classification performance while ensuring uniformity in the input features utilized, namely, BERT-VAD for text, MFCC, Delta, and Delta² for audio, and VGG16 for image.

The assessment uncovered a clear performance ranking among the three models. The Transformer-Based Fusion architecture demonstrated exceptional performance, achieving an Accuracy and F1-Score of 0.95. The performance exceeded that of the Late Fusion architecture, which achieved a score of 0.94, and the Early Fusion architecture, which recorded a score of 0.93. The Transformer-Based Fusion architecture was quantitatively proven to be the most effective strategy among the architectures or model types tested, highlighting the importance of integrating attention-based features for this task.

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

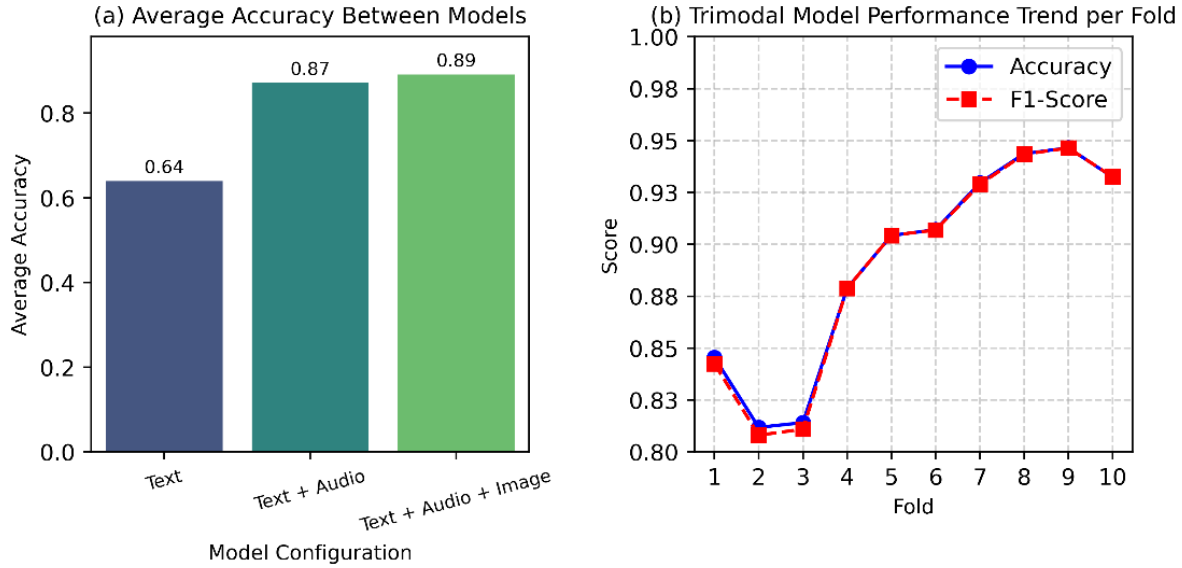

Figure 4. (a) Average Accuracy Between Models; (b) Trimodal Model Performance Trend per Fold

The Figure 4 presents a detailed overview of the proposed model's performance. The bar chart on the left illustrates a comparative ablation study, showcasing the average accuracy attained by three distinct model configurations: unimodal (Text), bimodal (Text + Audio), and the final trimodal (Text + Audio + Image) architecture. This chart illustrates a notable enhancement in performance as additional modalities are incorporated, with the trimodal model reaching the peak average accuracy of 0.89 or 89%. The line chart on the right illustrates the performance trend of the final trimodal model throughout the 10 folds of the cross-validation process. The plot displays the Accuracy represented by the blue line and the F1-Score indicated by the red line for each fold. The data indicates that although performance varies, which is a typical feature of cross-validation, the model reliably attains high scores (exceeding 0.90) starting from fold 5. The strong correlation between the Accuracy and F1-Score metrics suggests a balanced classification performance across various emotion labels, with the model demonstrating its highest effectiveness in the later folds.

4. DISCUSSIONS

An extensive analysis was conducted on the performance of the multimodal emotion classification model, which incorporates features from three primary modalities namely text using BERT and NRC-VAD, audio using MFCC, Delta-MFCC, and Delta^2-MFCC, and images using VGG16. The evaluation utilized the Stratified K-Fold Cross Validation method (k = 10) and achieved an accuracy of 95% on the ninth fold, indicating a significant proficiency in emotion recognition from multimodal input.

Testing each mode individually revealed that the text modality had the lowest accuracy relative to the other two modalities. This is probably attributable to the restricted context of emotions that may be conveyed just through writing. Conversely, the image modality yields the highest accuracy, as facial expressions visually convey emotions most directly. The auditory modality occupies an intermediary position, demonstrating a strong capacity to identify high-intensity emotions.

The integration of the three modalities using the Transformer-Based Fusion architecture technique yields a substantial enhancement in accuracy. This combination effectively harnesses the distinct advantages of each modality synergistically, particularly in the classification of ambiguous emotions like "fear" and "surprise," which frequently overlap both visually and verbally.

These research are juxtaposed with other prior research that share a comparable focus. For instance, in the research conducted by Khan et al. [7], which focused on emotion categorization utilizing

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

two modalities namely text and visual, a maximum accuracy of 73.55% was achieved through a late fusion method employing the Question, Visual, and Subtitles datasets. The findings suggest that including audio modalities and employing transformer-based fusion topologies, as demonstrated in this study, can yield substantial enhancements in performance.

This evaluation demonstrates that the integration of text, audio, and image-based modalities using a Transformer-Based Fusion approach yields substantial accuracy enhancements and exhibits flexibility in addressing the variability of emotional expressions in complex inputs. This methodology is very applicable to numerous real-world contexts, including emotion analysis in customer service systems, human-computer interaction, and psychological well-being assessment. It is important to acknowledge that augmenting model complexity leads to elevated computing demands. Consequently, the deployment of this system on resource-constrained devices necessitates adjustments via model optimization or the application of effective model compression techniques.

5. CONCLUSION

The multimodal emotion classification system, utilizing a Transformer-Based Fusion architecture, achieved an accuracy of 95% in the ninth k-fold validation based on the test data. The integration of features from three modalities namely text using BERT and NRC-VAD, audio using MFCC, Delta-MFCC, and Delta^2-MFCC, and pictures using VGG16, demonstrated a considerable enhancement in performance relative to their individual applications. The test results from the ninth fold shown elevated and consistent precision, recall, and F1-score metrics across all emotion categories. The “neutral” emotion achieved the highest f1-score of 0.98, succeeded by “happy” and “sad” at 0.97, “surprised” and “fear” at 0.92, and “angry” recorded the lowest F1-score at 0.91, yet remained within a high range. The mean macro and weighted F1-scores were both 0.95, showing the model's consistency across classes.

Analysis by modality indicates that the image modality yields the highest accuracy relative to the other two, as facial expressions more effectively convey emotions. The text modality has the lowest accuracy due to the contextual and ambiguous nature of emotional expressions in written form, but the audio modality ranks in the middle and effectively differentiates high-intensity emotions.

Furthermore, the implementation of the SMOTE oversampling technique facilitates the equilibrium of class distribution without inducing overfitting, as it is executed with a maximum ratio of 1:2. This enhances the stability of classification outcomes among classes and augments performance in classes with less data.

This study demonstrates that the multimodal Transformer-Based Fusion methodology possesses significant potential for use in many real-world scenarios, including user emotion analysis, AI-driven psychological services, and enhanced empathetic and adaptable human-computer interactions.

For future development, it is advisable to evaluate the system using data from diverse real-world contexts, including noise, picture distortion, or non-standard text, and to compare alternative fusion architecture such as cross-modal attention or joint embedding. Evaluating devices with constrained resources is crucial for assessing the system's efficiency in practical applications.

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 authors express their sincere gratitude to all parties who have provided valuable support and contributions, enabling the successful completion of this journal.

REFERENCES

[1] World Health Organization, "World Mental Health Report: Transforming mental health for all," 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240049338

[2] C. Singla, S. Singh, P. Sharma, N. Mittal, and F. Gared, "Emotion recognition for human–computer interaction using high-level descriptors," *Sci Rep*, vol. 14, no. 1, p. 12122, May 2024, doi: 10.1038/s41598-024-59294-y.

[3] R. Zhen, W. Song, Q. He, J. Cao, L. Shi, and J. Luo, "Human-Computer Interaction System: A Survey of Talking-Head Generation," *Electronics (Basel)*, vol. 12, no. 1, p. 218, Jan. 2023, doi: 10.3390/electronics12010218.

[4] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Syst Appl*, vol. 237, p. 121692, Mar. 2024, doi: 10.1016/j.eswa.2023.121692.

[5] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowl Based Syst*, vol. 244, p. 108580, May 2022, doi: 10.1016/j.knosys.2022.108580.

[6] W. Chango, R. Cerezo, M. Sanchez-Santillan, R. Azevedo, and C. Romero, "Improving prediction of students' performance in intelligent tutoring systems using attribute selection and ensembles of different multimodal data sources," *J Comput High Educ*, vol. 33, no. 3, pp. 614–634, Dec. 2021, doi: 10.1007/s12528-021-09298-8.

[7] A. U. Khan, A. Mazaheri, N. da V. Lobo, and M. Shah, "MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering," *ArXiv*, Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.14095

[8] B. Pan, K. Hirota, Z. Jia, and Y. Dai, "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods," *Neurocomputing*, vol. 561, p. 126866, Dec. 2023, doi: 10.1016/j.neucom.2023.126866.

[9] D. Salvi, B. Hosler, P. Bestagini, M. C. Stamm, and S. Tubaro, "TIMIT-TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection," *IEEE Access*, vol. 11, pp. 50851–50866, 2023, doi: 10.1109/ACCESS.2023.3276480.

[10] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowl Based Syst*, vol. 244, p. 108580, May 2022, doi: 10.1016/j.knosys.2022.108580.

[11] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *J Big Data*, vol. 9, no. 1, p. 15, Dec. 2022, doi: 10.1186/s40537-022-00564-9.

[12] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works," *Trans Assoc Comput Linguist*, vol. 8, pp. 842–866, Dec. 2020, doi: 10.1162/tacl_a_00349.

[13] A. Bello, S.-C. Ng, and M.-F. Leung, "A BERT Framework to Sentiment Analysis of Tweets," *Sensors*, vol. 23, no. 1, p. 506, Jan. 2023, doi: 10.3390/s23010506.

[14] K. Yang, T. Zhang, H. Alhuzali, and S. Ananiadou, "Cluster-Level Contrastive Learning for Emotion Recognition in Conversations," *IEEE Trans Affect Comput*, vol. 14, no. 4, pp. 3269–3280, Oct. 2023, doi: 10.1109/TAFFC.2023.3243463.

[15] S. P. Mishra, P. Warule, and S. Deb, "Speech emotion recognition using MFCC-based entropy feature," *Signal Image Video Process*, vol. 18, no. 1, pp. 153–161, Feb. 2024, doi: 10.1007/s11760-023-02716-7.

[16] F. S. AL-ANZI, "IMPROVED NOISE-RESILIENT ISOLATED WORDS SPEECH RECOGNITION USING PIECEWISE DIFFERENTIATION," *Fractals*, vol. 30, no. 08, Dec. 2022, doi: 10.1142/S0218348X22402277.

[17] A. Jawale and G. Magar, "MFCC Delta–Delta Energy Feature Extraction for Clustering of Road Surface Types," *International Journal of Pavement Research and Technology*, vol. 16, no. 3, pp. 631–646, May 2023, doi: 10.1007/s42947-022-00153-2.

[18] F. S. AL-ANZI, "IMPROVED NOISE-RESILIENT ISOLATED WORDS SPEECH RECOGNITION USING PIECEWISE DIFFERENTIATION," *Fractals*, vol. 30, no. 08, Dec. 2022, doi: 10.1142/S0218348X22402277.

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

[19] A. Alshehri and D. AlSaeed, "Breast Cancer Diagnosis in Thermography Using Pre-Trained VGG16 with Deep Attention Mechanisms," *Symmetry (Basel)*, vol. 15, no. 3, p. 582, Feb. 2023, doi: 10.3390/sym15030582.

[20] W. Bakasa and S. Viriri, "VGG16 Feature Extractor with Extreme Gradient Boost Classifier for Pancreas Cancer Prediction," *J Imaging*, vol. 9, no. 7, p. 138, Jul. 2023, doi: 10.3390/jimaging9070138.

[21] T. M. Saravanan, K. Karthiha, R. Kavinkumar, S. Gokul, and J. P. Mishra, "A novel machine learning scheme for face mask detection using pretrained convolutional neural network," *Mater Today Proc*, vol. 58, pp. 150–156, 2022, doi: 10.1016/j.matpr.2022.01.165.

[22] J. Zhang, A. Liu, D. Wang, Y. Liu, Z. J. Wang, and X. Chen, "Transformer-Based End-to-End Anatomical and Functional Image Fusion," *IEEE Trans Instrum Meas*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3200426.

[23] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020, doi: 10.1109/ACCESS.2020.3026823.

[24] Y. Wang, Y. Gu, Y. Yin, Y. Han, H. Zhang, S. Wang *et al.*, "Multimodal transformer augmented fusion for speech emotion recognition," *Front Neurorobot*, vol. 17, May 2023, doi: 10.3389/fnbot.2023.1181598.

[25] M. T R, V. K. V, D. K. V, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, p. 100247, Dec. 2023, doi: 10.1016/j.health.2023.100247.

[26] A. K. Adepu, S. Sahayam, U. Jayaraman, and R. Arramraju, "Melanoma classification from dermatoscopy images using knowledge distillation for highly imbalanced data," *Comput Biol Med*, vol. 154, p. 106571, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106571.

[27] H. Abu-Nowar, A. Sait, T. Al-Hadhrami, M. Al-Sarem, and S. Noman Qasem, "SENSES-ASD: a social-emotional nurturing and skill enhancement system for autism spectrum disorder," *PeerJ Comput Sci*, vol. 10, p. e1792, Feb. 2024, doi: 10.7717/peerj-cs.1792.

[28] G. G. Pushpa, J. Kotti, and Ch. Bindumadhuri, "Face Emotion Recognition Based on Images Using the Haar-Cascade Front End Approach," 2024, pp. 331–339. doi: 10.1007/978-3-031-48888-7_28.

[29] G. P. Kusuma, J. Jonathan, and A. P. Lim, "Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 6, pp. 315–322, 2020, doi: 10.25046/aj050638.

[30] S. Mohammad, "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 174–184. doi: 10.18653/v1/P18-1017.

[31] M. Hou, Z. Zhang, C. Liu, and G. Lu, "Semantic Alignment Network for Multi-Modal Emotion Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 5318–5329, Sep. 2023, doi: 10.1109/TCSVT.2023.3247822.

[32] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal Multi-label Emotion Detection with Modality and Label Dependence," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 3584–3593. doi: 10.18653/v1/2020.emnlp-main.291.

[33] A. Rahman Yusuf and A. Prasetyo, "The Use of Information Retrieval in Student Academic Document Plagiarism Detection System," *bit-Tech*, vol. 6, no. 2, pp. 235–240, Dec. 2023, doi: 10.32877/bt.v6i2.1063.

[34] H. A. Owida, A. Al-Ghraibah, and M. Altayeb, "Classification of Chest X-Ray Images using Wavelet and MFCC Features and Support Vector Machine Classifier," *Engineering, Technology & Applied Science Research*, vol. 11, no. 4, pp. 7296–7301, Aug. 2021, doi: 10.48084/etasr.4123.

Vol. 6, No. 4, Agustus 2025, Page. 2586-2601
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4915

[35] F. Aldi, F. Hadi, N. A. Rahmi, and S. Defit, "Standardscaler's Potential in Enhancing Breast Cancer Accuracy Using Machine Learning," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 5, no. 1, pp. 401–413, Dec. 2023, doi: 10.37385/jaets.v5i1.3080.

[36] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Applied Intelligence*, vol. 52, no. 5, pp. 5457–5474, Mar. 2022, doi: 10.1007/s10489-021-02524-x.

[37] T. Hoang, T.-T. Do, T. V. Nguyen, and N.-M. Cheung, "Multimodal Mutual Information Maximization: A Novel Approach for Unsupervised Deep Cross-Modal Hashing," *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 9, pp. 6289–6302, Sep. 2023, doi: 10.1109/TNNLS.2021.3135420.

[38] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[39] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020, doi: 10.1109/ACCESS.2020.3026823.

[40] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning," *IEEE Access*, vol. 11, pp. 14742–14751, 2023, doi: 10.1109/ACCESS.2023.3244390.

[41] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers in Nanotechnology*, vol. 4, Aug. 2022, doi: 10.3389/fnano.2022.972421.

[42] R. Romijnders, E. Warmerdam, C. Hansen, J. Welzel, G. Schmidt, and W. Maetzler, "Validation of IMU-based gait event detection during curved walking and turning in older adults and Parkinson's Disease patients," *J Neuroeng Rehabil*, vol. 18, no. 1, p. 28, Dec. 2021, doi: 10.1186/s12984-021-00828-0.