Parallelizing CNN and Transformer Encoders for Audio Based Emotion Recognition in English Language

Adam Satria Adidarma*, Kelly Rossa Sungkono†, and Shintami Chusnul Hidayati‡
* Sepuluh Nopember Institute of Technology, Indonesia
E-mail: adam.19051@mhs.its.ac.id
† Sepuluh Nopember Institute of Technology, Indonesia
E-mail: kelly@its.ac.id
‡ Sepuluh Nopember Institute of Technology, Indonesia
E-mail: shintami@its.ac.id

Abstract—**Artificial intelligence (AI) has greatly impacted diverse industries, witnessing a 37% adoption increase from 2018 to 2019. Within AI, speech emotion recognition (SER) focuses on identifying emotions in speech. Emotions play a crucial role in human communication and have been the subject of increasing research in recent years. While current research emphasizes visual indicators, emotion is a multimodal concept, requiring the study of visual, tactile, vocal, and physiological cues. This paper proposes an SER system using a parallelized CNN model with a Transformer Encoder Block to classify emotions, including anger, disgust, fear, happy, neutral, and sad. Evaluation is performed on publicly available English audio emotion datasets such as CREMA-D, RAVDESS, and SAVEE. The Mel Frequency Cepstrum Coefficients (MFCC) were used to extract the features of these sounds. The Transformer Encoder and CNN model achieves the best overall performance across datasets with 74% average accuracy. The Transformer Encoder Block excels in processing long sequences, especially in natural language processing, while CNNs capture local patterns, ideal for analyzing spectrogram images from audio signals.**

I. Introduction

The widespread influence of Artificial Intelligence (AI) has had a profound impact on various industries and sectors of society. The adoption of AI has grown significantly in recent years, as evidenced by a 37% increase during 2018-2019 [1]. Speech Emotion Recognition (SER) has emerged as a promising application in the context of AI. SER involves the task of autonomously recognizing the emotional aspects of speech, independent of its semantic content. While humans possess the natural ability to perform this task, achieving automated recognition using programmable devices remains an active area of research [2]. The concept of humans interacting with computers as if they were other people, as discovered in studies of human-computer interaction [3], highlights the importance of emotions and their connection to machines. Emotionally intelligent machines can significantly enhance user experience and overall machine performance [4]. Emotions play a crucial role in human communication, leading to an increase in research efforts aimed at understanding human emotions [5]. Emotion classification using speech has found application in various fields, such as security systems, psychology, computer vision, and interactive computer designs. While existing studies have primarily focused on visual modalities for emotion detection, it is important to consider the multimodal nature of emotions, requiring interdisciplinary investigations involving visual cues, tactile communication, vocalization, and physiological indications [6].

Recent research has shown considerable interest in using multiple Support Vector Machine (SVM) methods, both linear and nonlinear, for emotion classification and sound detection [7]. Some studies have explored the effectiveness of transfer learning with pre-trained deep learning models to improve accuracy. Notably, deep learning-based Convolutional Neural Network (CNN) methods have outperformed handcrafted feature-based SVM methods in image classification [8]. Deep learning models, such as GPT-3 and BERT, have also gained prominence, especially in the field of Natural Language Processing (NLP), due to their self-attention mechanism in the Transformer Network [9][10]. This paper aims to implement a LeNet-based CNN architecture in conjunction with a parallel Transformer Encoder Block, employing a self-attention mechanism, to detect human emotions using an English audio dataset. This model will be compared against a Support Vector Machine (SVM) and a deep learning convolution-based architecture, which includes the LetNet-based CNN, and the Convolutional Recurrent Neural Network (CRNN) model. This comparative analysis will evaluate the performance and effectiveness of the Parallel Transformer Encoder with the CNN model in comparison to these alternative machine-learning approaches. The objective is to identify the most suitable and accurate method for emotion detection, considering factors such as expressive feature representation and the ability to predict different emotions based on the overall structure of the Mel Frequency Cepstrum Coefficients (MFCC) plot.

II. Related Works

In the context of detecting human emotions from audio data, the selection and extraction of audio features are crucial in the detection of human emotions through audio data. Recent studies have utilized the Sequential Minimal Optimization (SMO) algorithm to analyze sounds and train SVM models. Emotional characteristics such as arousal (the level of autonomic activation) and valence (the level of pleasantness) are observed to understand human emotion [11]. Various research works have employed different methods for audio feature extraction, including prosodic, excitation, vocal tract, spectral descriptors, and MFCC in combination with the SMO algorithm [12]. To facilitate emotion recognition from speech, comprehensive collections of emotional speech data have been developed. These collections consist of lab recordings, which are high-quality recordings made in controlled environments with linguistic experts, and non-lab recordings, which capture emotions in natural scenarios. Notable emotional speech corpora include EmoDB [13], IEMOCAP [14], and AESDD [15] for lab recordings and DAPS [16], Freefield1010 [17], and CHEAVD [18] for non-lab recordings. These corpora provide a diverse range of emotional speech data for research purposes.

Recent studies have explored various deep-learning architectures for speech representation. The wav2vec [19] model introduced unsupervised learning for speech recognition, while the HuBERT [20] model addressed issues with self-supervised learning by applying prediction loss to masked regions. Other models such as UniSpeech [21] and WavLM [22] have further improved speech recognition performance by leveraging both supervised and unsupervised data. In the field of emotion classification, popular methods include SVM [7], Hidden Markov Model (HMM)[23], and Recurrent Neural Network (RNN) [24]. These methods, along with feature extraction and emotional output recognition, are integral components of a speech-emotion recognition system. The classification models in this paper were trained using datasets including CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset) [25], RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [26], and SAVEE (Surrey Audio-Visual Expressed Emotion) [27]. These datasets contain recordings of actors expressing various emotions, providing valuable resources for studying acoustic features associated with different emotional expressions.

III. Methodology

The structure of the model includes two main sections running in parallel: a CNN block and a Transformer encoder block. Fig. 1 provides a diagram that visually represents the connections and information flow within the architecture. This diagram helps to grasp the complex operations of the model and facilitates comprehension of the following sections.

The CNN architecture utilized in this paper leverages recent advancements in image and sequence processing. It consists of convolutional and pooling layers inspired by the classic LeNet architecture, which extract features from the input data and downsample the feature maps. The extracted features are then processed by fully-connected layers, leading to the network's final output. To obtain a probability distribution over possible classes, the SoftMax function is applied. Despite its simplicity compared to modern CNNs, LeNet has demonstrated success in various classification tasks, such as handwritten digit recognition.

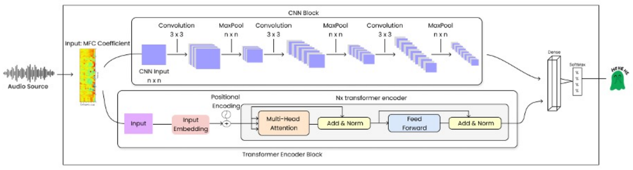

Fig. 1. Model Architecture.

For the Transformer architecture, the same design was used as described by [28]. This paper utilized the encoder blocks, which are crucial components of the Transformer architecture. These encoder blocks process the input sequence, extract relevant features, and pass them to the fully-connected layers for further processing, ultimately producing the network's final output. The fully-connected layers map the extracted features to the desired output format, such as classification probabilities or predicted values. Each encoder block comprises a self-attention layer followed by a feedforward layer. The self-attention layer employs a dot-product attention mechanism to calculate attention weights between input elements. These weights are then utilized to compute a weighted sum of the input elements, generating the output of the self-attention layer. The feedforward layer consists of two linear transformations separated by a ReLU activation function. It takes the output of the self-attention layer as input and produces the final output of the encoder block.

The parallel architecture was used in order to enable the network to concurrently process multiple features by employing a series of inception modules. These modules are concatenated and fed into a fully-connected (dense) layer. The parallel architecture has demonstrated good performance while being efficient in terms of parameters and computation time.

A. Dataset

The experiment utilizes three distinct English audio datasets, namely CREMA-D, RAVDESS, and SAVEE. These datasets have been widely embraced in the field of audio emotion recognition and serve as valuable resources for training and evaluating models. Each dataset contains a unique collection of recordings featuring emotional speech, enabling a comprehensive analysis and comprehension of various facets of emotional expression in audio data. The utilization of multiple datasets enables the evaluation of model performance across diverse contexts and ensures a more representative analysis of audio emotion recognition capabilities.

1) CREMA-D: The dataset was created as a valuable resource for audio-visual scene understanding research. It encompasses a substantial collection of 7,442 audio and video clips capturing everyday scenarios, including conversations, laughter, and singing in diverse environments. The dataset features recordings from 91 actors, with a balanced representation of 48 male and 43 female individuals, spanning ages 20 to 74 and encompassing various racial and ethnic backgrounds (African American, Asian, Caucasian, Hispanic, and Unspecified). The actors delivered 12 sentences, expressing six distinct emotions (anger, disgust, fear, happy, neutral, and sad) across four emotion levels (low, medium, high, and unspecified). Recordings took place in a wide range of settings, such as homes, offices, parks, and streets. The dataset offers considerable diversity in terms of age, gender, and ethnicity among the individuals portrayed, enabling effective training of models that generalize well in real-world scenarios. Detailed annotations accompany each clip, providing comprehensive information about the audio and visual content, as well as the people and objects present within the scenes. These annotations empower researchers to leverage the dataset for various tasks, including speech recognition, object detection, and facial recognition.

2) RAVDESS: This dataset is a publicly available resource comprising emotional speech recordings. It consists of 1,470 recordings performed by 24 professional actors, ensuring an equal gender distribution with 50% male and 50% female actors. These actors were instructed to convey various emotions, including calm, happy, sad, angry, fear, and surprised. Notably, the RAVDESS dataset's notable strength lies in its balanced representation of male and female voices, addressing the importance of gender differences in emotion perception and expression. By incorporating an equal number of actors from both genders, the dataset provides a more representative sample of emotional expressions across genders. Furthermore, the RAVDESS dataset also includes recordings of speech in both neutral and accented English. This addition is significant as accents can influence the perception and expression of emotions. By including accented English recordings, the RAVDESS dataset offers a more diverse range of emotional expressions, better reflecting real-world scenarios where emotions are expressed across different accents and cultures.

3) SAVEE: A widely recognized and extensively employed audio dataset that contains 480 recordings of British English sentences, each lasting approximately 4-5 seconds. It comprises recordings from four male speakers, with each speaker conveying seven distinct emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The sentences were carefully selected from the standard TIMIT corpus and phonetically balanced for each specific emotion. The primary objective behind creating this dataset was to support advancements in the field of speech emotion recognition, and it has been widely adopted in numerous research studies within this domain. The recordings took place in a visual media lab, utilizing high-quality audio-visual equipment. They were meticulously processed and labelled, resulting in a dataset of exceptional quality, well-suited for training and evaluating emotion recognition models. The dataset serves as a valuable resource, offering a comprehensive range of emotional expressions and providing researchers with a reliable foundation for developing and assessing state-of-the-art models in the field of speech emotion recognition.

B. Model Implementation

This section describes the steps involved in conducting the research and explains the progression of the study from start to finish. The various stages of the research are depicted in Fig. 1, which provides a visual representation of the Parallel Transformer Encoder with CNN model architecture flow.

1) Pre-processing: The initial step involves preprocessing the audio data obtained from the datasets. Preprocessing plays a crucial role in analyzing audio data as it entails converting raw audio signals into a format suitable for extracting meaningful information using machine learning algorithms. Audio data is often intricate, consisting of a wide range of frequencies, background noise, and other variations that can pose challenges in identifying and extracting relevant features. To address these challenges, preprocessing incorporates various steps aimed at cleaning and transforming the raw audio data into a format that is more conducive to analysis. In the context of audio emotion recognition, preprocessing assumes particular significance. It entails transforming the raw audio signals into a format that facilitates training a machine-learning model to recognize different emotions based on the acoustic characteristics of speech. This preprocessing involves a series of steps, such as downsampling the audio to a customized target sample rate, truncating the audio to a specific duration, and eliminating any silence preceding the actors' speech. These steps are essential to ensure consistency in the audio data and eliminate irrelevant noise or silence that may impact the accuracy of the emotion recognition model.

To standardize the sampling frequency of the audio files and align it with the default sample rate used by most deep learning frameworks, a custom sample rate was chosen. Furthermore, an offset was applied to the audio files to remove any pre-speech silence, ensuring that only the emotional speech segments were retained for analysis. This step aimed to eliminate unnecessary background noise or silence that could interfere with the emotion recognition process. Subsequently, the audio files were loaded and evenly divided into training and testing sets across different emotions. The training set was used to train the emotion recognition model, while the testing set was used to evaluate the model's performance on unseen data. Random splitting was employed to ensure the training and testing sets represented the entire dataset and to avoid any bias in the model's performance. The emotion labels will be mapped numerically into six different categories for training using a mapping of 'angry': 0, 'fear': 1, 'disgust': 2, 'happy': 3, 'neutral': 4, 'sad': 5. This mapping was performed to convert the categorical labels into a format suitable for training and evaluating the machine-learning model. To provide a clear depiction of the preprocessing process, Fig. 2 presents a flowchart diagram illustrating the sequence of operations involved.

2) Feature Extraction: The subsequent step involves extracting audio features from the input data using the Mel-Frequency Cepstral Coefficients (MFCCs). MFCC is a widely used technique in speech and audio processing for its ability

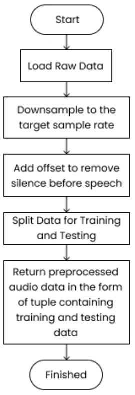
Start
Load Raw Data
Downsample to the target sample rate
Add offset to remove silence before speech
Split Data for Training and Testing
Return preprocessed audio data in the form of tuple containing training and testing data
Finished

Fig. 2. Pre-processing Flowchart Diagram.

to compactly and efficiently capture the spectral characteristics of an audio signal. It derives MFCCs from the power spectrum of the audio signal, utilizing the Mel scale, a non-linear scale based on the perceived frequency of sound by the human ear. This characteristic makes MFCCs particularly suitable for tasks like speech recognition and speaker identification, where human auditory perception is paramount. Fig. 3 illustrates the workflow for extracting MFCC features. Once the MFCC features are extracted, they can serve as inputs for machine learning models in speech emotion recognition tasks, enabling the analysis and classification of emotions conveyed through audio signals.

The implementation employs the librosa library to retrieve the audio signal from a file and extract the MFCCs through a series of sequential stages. The first step involves computing the Short-Time Fourier Transform (STFT) of the audio waveform using a specified window size and hop length. This process divides the audio signal into short frames of equal length, with the hop length typically being half the frame size. Each audio frame is typically windowed using the Hanning function, which minimizes spectral leakage and enhances the frequency resolution of the STFT. The subsequent step calculates the magnitude spectrogram of the STFT. The magnitude spectrogram is then passed through a Mel filter bank, approximating the frequency response of the human auditory system. Following this, the resulting Mel spectrogram is converted into decibel (dB) units on a logarithmic scale. This conversion compresses the magnitude values and provides a more accurate representation of the relative loudness in each frequency bin. Finally, the MFCC coefficients are computed by performing a discrete cosine transform of the log-mel spectrogram. This step transforms the Mel spectrogram into a sequence of cepstral coefficients that characterize the spectral envelope of the audio signal.

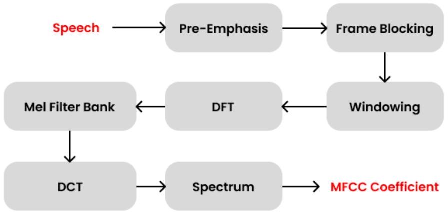

Fig. 3. MFCC Block Diagram.

3) Model Architecture Design: Once the audio features have been extracted, the final step is to design the model architecture and utilize the extracted features as input for classifying human emotions. The Parallel Transformer Encoder with CNN Architecture consists of two blocks: the CNN block and the Transformer block, working in parallel. The CNN block provides a spatial feature representation of the input data, while the Transformer block focuses on accurately modelling the temporal relationships between pitch transitions in human emotions. The expansion of CNN filter channels and reduction of feature maps aim to achieve expressive feature representation with minimal computational cost. The Transformer encoder learns to predict frequency distributions of different emotions based on the global structure of the MFCC plot for each emotion.

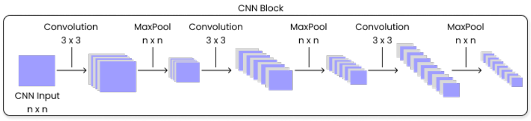

Fig. 4. CNN Block.

The CNN block depicted in Fig. 4 starts with the input layer that receives audio features with a specific size and number of channels. The convolution layer applies filters to the input data through convolution, generating a set of feature maps. Batch normalization normalizes the output from the convolution layer, enhancing the stability and efficiency of the model. The activation layer applies a non-linear activation function to introduce complexity in the model and capture intricate relationships in the data. The pooling layer downsamples the feature maps using a pooling operation, reducing their size and computational complexity. This sequence of convolutional, batch normalization, activation, and pooling layers is repeated three times in the model architecture.

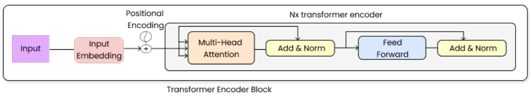

Fig. 5. Transformer Encoder Block.

The implementation of the Transformer encoder shown in Fig. 5, is designed for processing sequential audio data. It comprises self-attention layers and feedforward layers, which predict frequency distributions of different emotions based on the overall structure of the MFCCs for each emotion. The implementation initializes an empty list for the output sequence. The input sequence is embedded by applying an embedding matrix to each element, generating a sequence of embedded vectors. These embedded vectors are passed through self-attention layers to compute context-aware representations. Each self-attention layer employs the attention mechanism to compute a weighted sum of the input vectors, considering their relationships. The resulting context-aware representations go through feedforward layers to obtain transformed representations. Each feedforward layer applies a linear transformation followed by a non-linear activation function. The transformed representations are then used to compute the output sequence by applying a linear transformation and an activation function to each element.

The outputs from the CNN and Transformer models are concatenated in the last layer of the Parallel Transformer Encoder with the CNN Architecture model. The resulting tensor is passed through a dense layer with a softmax activation function for prediction. The softmax function converts the arbitrary real-valued scores into a probability distribution over the classes, enabling the classification of emotions. The dense layer implementation begins by combining the output tensors of the CNN and Transformer models, which have the shape (batch_size, feature_size) for each model, respectively. The combined output is then passed through a linear layer with the number of class units, representing the six different emotional states in the dataset. The weights and biases of the dense layer are learned during training to transform the combined output into the final prediction. Finally, the softmax activation function converts the prediction scores into a probability distribution over the classes. The class with the highest probability is considered the model's prediction. Table I provides the layer architecture overview of the Parallel Transformer Encoder with CNN Architecture model.

C. Model Evaluation

Model evaluation is a step in assessing the performance and suitability of deep learning models for specific tasks. In the context of speech emotion recognition, this section provides an overview of the considerations for evaluating deep learning models. There are several ways to evaluate the performance of deep learning models. This paper compares the performance of four different machine learning models on a speech emotion recognition task. These models include the standard machine learning SVM model, a LeNet-based CNN architecture model, the CRNN model, and the Parallel Transformer Encoder with CNN Architecture. A combination of different evaluation metrics will be used to evaluate the performance of these models, including:

1) Training and Validation Accuracy: The training and validation accuracy of each model were monitored to identify and address potential issues such as overfitting and underfitting. Tracking the validation accuracy assists in selecting optimal hyperparameter values for improved model performance. This approach will show how effectively the models learn the classification task and identify any optimization-related challenges.

TABLE I
MODEL ARCHITECTURE LAYER.

Block	Layer	Output Shape	Parameter
1	Conv2D Layer 1	[16, 40, 80]	160
	ELU	[16, 40, 80]	-
	Batch Normmalization	[16, 40, 80]	32
	MaxPool	[16, 20, 40]	-
	Dropout	[16, 20, 40]	-
	Conv2D Layer 2	[32, 20, 40]	4,640
	ELU	[32, 20, 40]	-
	Batch Normmalization	[32, 20, 40]	64
	MaxPool	[32, 5, 10]	-
	Dropout	[32, 5, 10]	-
	Conv2D Layer 3	[64, 5, 10]	18,496
	ELU	[64, 5, 10]	-
	Batch Normmalization	[64, 5, 10]	128
	MaxPool	[64, 1, 2]	-
	Dropout	[64, 1, 2]	-
2	Conv2D Layer 1	[16, 40, 80]	160
	ELU	[16, 40, 80]	-
	Batch Normmalization	[16, 40, 80]	32
	MaxPool	[16, 20, 40]	-
	Dropout	[16, 20, 40]	-
	Conv2D Layer 2	[32, 20, 40]	4,640
	ELU	[32, 20, 40]	-
	Batch Normmalization	[32, 20, 40]	64
	MaxPool	[32, 5, 10]	-
	Dropout	[32, 5, 10]	-
	Conv2D Layer 3	[64, 5, 10]	18,496
	ELU	[64, 5, 10]	-
	Batch Normmalization	[64, 5, 10]	128
	MaxPool	[64, 1, 2]	-
	Dropout	[64, 1, 2]	-
3	Transformer Encoder Layer 1 – 6	[2, 40]	48,232
	FC Layer 1	[6]	1,782
	Softmax	[6]	-

* '-' indicates that there are no specific parameters associated with the layer or operation.

2) Test Set Accuracy: This metric offers a reliable summary of model performance on data that was not encountered during training. Evaluating test set accuracy enables a realistic assessment of the model's generalization ability, essential for gauging its suitability for deployment and real-world performance.

3) Confusion Matrix: A confusion matrix presents the number of true positive, true negative, false positive, and false negative predictions made by the model. It facilitates an assessment of model performance across different emotion classes, offering a detailed understanding of their strengths and weaknesses.

4) Precision, Recall, and F1 Score: These metrics are commonly employed in assessing the performance of deep learning models, particularly for classification tasks. They enable a comprehensive understanding of model performance and comparison across the six different emotion classes in the tested dataset.

5) Weighted Average of F1 Scores: The weighted average of F1 scores addressed imbalanced classes for a comprehensive model performance evaluation.

IV. RESULTS & DISCUSSION

The Parallel Transformer Encoder with CNN model architecture combines the strengths of the Transformer Encoder and

CNN to effectively process sequential data like speech signals. The architecture leverages the global structure of MFCC plots associated with each emotion, allowing the model to learn frequency distributions and capture temporal patterns. This comprehensive approach enhances emotion classification performance by considering both local and global dependencies.

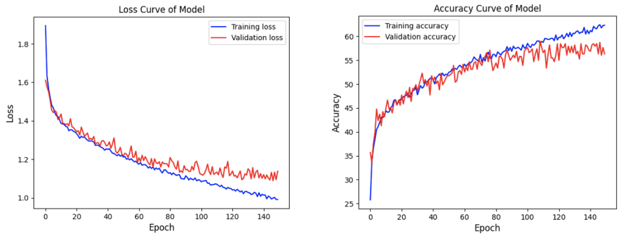

Fig. 6. CREMA-D Loss & Accuracy Curve.

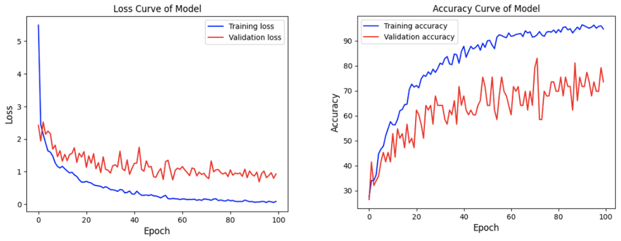

Fig. 7. CREMA-D Loss & Accuracy Curve.

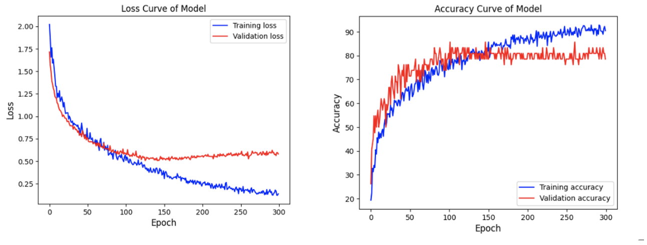

Fig. 8. CREMA-D Loss & Accuracy Curve.

The model was trained on the CREMA-D, RAVDESS, and SAVEE datasets, and the training process aimed to optimize the loss function and achieve high accuracy on the validation set, as illustrated in Fig. 6, Fig. 7, and Fig. 8 corresponding to each dataset. The loss and accuracy curves demonstrated the model's performance varied across datasets, with different convergence points and peak accuracies. The model achieved the highest accuracy of 59% on the CREMA-D dataset, 79% on the RAVDESS dataset, and 83% on the SAVEE dataset.

The performance evaluation of the model was analyzed using confusion matrices presented in Table II, Table III, and Table IV for the CREMA-D, RAVDESS, and SAVEE datasets, respectively. These matrices provided insights into the model's performance and its ability to accurately classify emotions. The results demonstrated varying levels of confusion observed across the different datasets. Specifically, the model showed high levels of confusion between sad and neutral for the CREMA-D and RAVDESS datasets. Additionally, the happy emotion displayed the lowest true positive rate, indicating instances where the model misclassified happy samples into other emotion classes. For the SAVEE dataset, confusion was observed between fear and sad emotions. These results emphasize the model's performance variations depending on the dataset being evaluated. The observed levels of confusion between emotion classes in the confusion matrices highlight both the strengths and weaknesses of the model for each tested dataset.

TABLE II
CREMA-D Confusion Matrix

Class	Predicted Values					
	Angry	Fear	Disgust	Happy	Neutral	Sad
Angry	41	3	8	10	2	0
Fear	2	40	6	2	6	8
Disgust	4	3	34	3	10	10
Happy	4	9	9	31	9	2
Neutral	1	3	6	1	39	4
Sad	1	4	9	1	13	35

TABLE III
RAVDESS Confusion Matrix

Class	Predicted Values					
	Angry	Fear	Disgust	Happy	Neutral	Sad
Angry	8	0	1	0	0	0
Fear	0	8	0	0	0	1
Disgust	0	1	8	1	0	0
Happy	0	2	0	8	0	0
Neutral	0	0	0	0	4	1
Sad	0	1	0	0	3	6

TABLE IV
SAVEE Confusion Matrix

Class	Predicted Values					
	Angry	Fear	Disgust	Happy	Neutral	Sad
Angry	3	1	0	0	1	1
Fear	0	4	0	0	0	2
Disgust	0	0	6	0	0	0
Happy	1	0	0	5	0	0
Neutral	0	0	0	0	12	0
Sad	0	0	0	0	1	5

Additionally, besides utilizing the confusion matrix, precision, recall, and F1-score metrics were employed to evaluate the model's capability in accurately classifying emotions within the considered dataset. These metrics offer additional insights that complement the information provided by the confusion matrix, resulting in a comprehensive assessment of the model's performance in emotion classification. Table V presents a comprehensive analysis of the model's accuracy in emotion classification, focusing specifically on the F1-scores for each emotion category across the tested datasets.

In terms of individual emotion classification, the angry emotion had the highest f1 score in both the CREMA-D and RAVDESS datasets, with f1 scores of 0.70 and 0.94, respectively. However, in the SAVEE dataset, the angry emotion

TABLE V
F1-SCORE EMOTION CLASSIFICATION

Emotion	F1-Score		
	CREMA-D	RAVDESS	SAVEE
Angry	0.70	0.94	0.60
Fear	0.63	0.76	0.73
Disgust	0.50	0.84	1.00
Happy	0.55	0.84	0.91
Neutral	0.59	0.67	0.92
Sad	0.57	0.67	0.71

had a lower F1 score compared to the other datasets, with a score of 0.65. On the other hand, the sad emotion had the lowest f1 score for all tested datasets, with scores of 0.57, and 0.67 for CREMA-D and RAVDESS, respectively, indicating the difficulty in accurately recognizing this emotion for this model. These findings demonstrate the performance and limitations of the Parallel Transformer Encoder with CNN model architecture for emotion recognition in speech signals. The results emphasize the model's strengths in accurately classifying certain emotions while highlighting areas for improvement, particularly in distinguishing closely related emotions.

The experiment compared the performance of various machine learning models for human emotion recognition from audio data. The models evaluated were SVM, LeNet CNN, CRNN, and the Parallel Transformer Encoder with CNN Architecture. The evaluation was conducted on three different datasets: CREMA-D, RAVDESS, and SAVEE. The data was split into training, validation, and testing sets with a ratio of 90:5:5. While the data preprocessing steps varied for each model, the feature extraction steps were consistent across all models, involving MFCC and normalization. The results are summarized in Table VI, showing the highest accuracy achieved by each model on specific datasets and overall. The accuracy of the models varied significantly across the datasets. For example, the Transformer Encoder and CNN model performed exceptionally well on the SAVEE dataset, achieving 83% accuracy. However, on the CREMA-D dataset, the same model achieved only 59% accuracy. This pattern was observed across all models, highlighting the strong influence of dataset characteristics on model performance.

TABLE VI
MODEL COMPARISON

Model	Dataset	Accuracy	Average
SVM	CREMA-D RAVDESS SAVEE	54% 72% 78%	68%
LeNet	CREMA-D RAVDESS SAVEE	52% 64% 67%	61%
CRNN	CREMA-D RAVDESS SAVEE	55% 70% 71%	65%
T. Encoder and CNN	CREMA-D RAVDESS SAVEE	59% 79% 83%	74%

Among the models, the Parallel Transformer Encoder with CNN Architecture demonstrated the best overall performance, with an average accuracy of 73.66% across all datasets. The SVM model also performed well, achieving an average accuracy of 68% across all datasets. SVM is a classic machine learning algorithm known for its success in classification tasks, especially with high-dimensional data. In this study, the SVM model effectively identified patterns in the audio data corresponding to different emotions. On the other hand, the CRNN model and LeNet CNN model exhibited lower accuracy compared to the other models. Despite the utility of CRNNs in sequential data processing, the CRNN model achieved an average accuracy of 65.33%, while the LeNet CNN model achieved an average accuracy of 61%.

V. CONCLUSIONS

In this study, the integration of Transformer Encoder and CNN architectures proved to be effective for detecting human emotions from audio data. By combining the strengths of both architectures, the model captured global dependencies and local patterns in the audio features, leading to improved performance in emotion detection. The optimized model architecture consisted of a Transformer Encoder block, two CNN-based blocks with a series of convolutional and pooling layers, and a final dense layer. Various optimization techniques, including hyperparameter tuning and regularization methods, were explored to enhance the accuracy of the model. The experimental results demonstrated that careful selection and fine-tuning of hyperparameters, along with appropriate regularization techniques, improved model performance and robustness. By comparing the accuracies of different models, the Parallel Transformer Encoder with CNN Architecture achieved the highest accuracy of 73.66% across all datasets, followed by the SVM model. However, the CRNN and LeNet CNN models showed comparatively lower accuracies in detecting emotions from audio data.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to the individuals and institutions whose invaluable contributions made this research possible. First and foremost, I extend my sincere appreciation to my supervisors, Kelly Rossa Sungkono, S.Kom., M.Kom. and Shintami Chusnul Hidayati, S.Kom., M.Sc., PhD, for their guidance, support, and encouragement throughout this project. Additionally, I would like to thank the Sepuluh Nopember Institute of Technology for providing the resources and facilities necessary for conducting this research. Finally, I am grateful for the unwavering support of their families and friends throughout the research process. Without the collective efforts and assistance of these individuals and organizations, this research would not have been possible.

References

[1] K. Costello, *Gartner survey shows 37 percent of organizations have implemented ai in some form*, https://www.gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have, Jan. 2019.

[2] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science*, p. 14, 2020.

[3] Ivar, R. Byron, and N. Clifford, *The media equation: How people treat computers, television, and new media like real people and places*. Center for the Study of Language and Inf, 1996.

[4] R. Cowie, "Emotion recognition in human-computer interaction," *IEEE*, pp. 32–80, 2001.

[5] J. Maria, "Understanding human emotions," *Journal of Russian & East European Psychology*, pp. 09–25, 2012.

[6] J. Heredia, Y. Cardinale, I. Dongo, and J. Díaz-Amado, "A multi-modal visual emotion recognition method to instantiate an ontology," *16th International Conference on Software Technologies*, pp. 453–464, 2021.

[7] A. Sonawane, M. U. Inamdar, and K. B. Bhangale, "Sound based human emotion recognition using mfcc & multiple svm," *IEEE*, pp. 1–4, 2017.

[8] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *arXiv*, 2018.

[9] T. Brown Mann, B. Ryder, N. Subbiah, *et al.*, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020, pp. 1877–1901.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2019.

[11] F. Citron, M. A. Gray, H. Critchley, B. Weekes, and E. C. Ferstl, "Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework," *Neuropsychologia*, pp. 79–89, 2014.

[12] F. Eyben Scherer, K. R. Schuller, B. W. Sundberg, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, pp. 190–202, 2016.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, 2005, pp. 1517–1520.

[14] C. Busso, M. Bulut, C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, pp. 335–359, 2008.

[15] N. Vryzas, R. Kotsakis, A. Liatsou, C. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *Journal of the Audio Engineering Society. Audio Engineering Society*, pp. 457–467, 2018.

[16] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? - a dataset, insights, and challenges," *IEEE Signal Processing Letters*, 2015.

[17] D. Stowell and M. D. Plumbey, "An open dataset for research on audio field recording archives: Freefield1010," *arXiv*, 2013.

[18] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: A chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, pp. 913–924, 2017.

[19] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," *arXiv*, 2019.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 3451–3460, 2021.

[21] C. Wang Wu, Y. Qian, Y. Kumatani, *et al.*, "Unispeech: Unified speech representation learning with labeled and unlabeled data," *Proceedings of the 38th International Conference on Machine Learning*, pp. 10 937–10 947, 2021.

[22] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1505–1518, 2022.

[23] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," *Proceedings of International Symposium on Computer Vision - ISCV*, pp. 265–270, 1995.

[24] S. Chamishka, I. Madhavi, R. Nawaratne, *et al.*, "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling," *SpringerLink*, pp. 35 173–35 194, 2022.

[25] H. Cao Cooper, D. G. Keutmann, M. K. Gur, R. C. Nenkova, A. Verma, and Ragini, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, pp. 377–390, 2014.

[26] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS ONE*, p. 13, 2018.

[27] P. Jackson and S. Haq, *Surrey audio-visual expressed emotion (savee) database*, http://kahlan.eps.surrey.ac.uk/savee/, Apr. 2015.

[28] A. Vaswani Shazeer, N. Parmar, N. Uszkoreit, *et al.*, *Attention is All You Need*. Curran Associates, Inc., 2017.