Alexandria University
Faculty of Engineering
Computer and Systems Engineering Department

Speech Emotion Recognition

Youssif Khaled Ahmed	**21011655**
Esmail Mahmoud Hassan	**21010272**
Ahmed Ayman Ahmed	**21010048**

Course: CSE: Pattern Recognition
Instructors: Prof. Dr. Marwan Torki, Eng. Ismail El-Yamany

Abstract

This report presents a comprehensive analysis of Speech Emotion Recognition (SER) systems, focusing on feature extraction techniques and comparative performance of various CNN architectures. We implement both 1D and 2D convolutional networks for emotion classification using the CREMA dataset, exploring different activation functions (ReLU, SiLU, ELU) and learning rates. Our experiments demonstrate the effectiveness of combined feature spaces and provide insights into the most confusing emotion classes in speech recognition.

Contents

1 Introduction

Speech is the most natural way of expressing ourselves as humans. It is only natural then to extend this communication medium to computer applications. Speech emotion recognition (SER) systems are collections of methodologies that process and classify speech signals to detect embedded emotions. These systems have numerous applications in human-computer interaction, customer service analysis, mental health monitoring, and entertainment.

SER presents unique challenges because emotional expressions in speech vary significantly across individuals, cultures, and contexts. The acoustic features that convey emotion can be subtle and often overlap with other speech characteristics. This assignment explores different approaches to SER, focusing on:

- Extracting relevant features from speech signals using both time and frequency domain representations
- Developing and comparing 1D and 2D CNN architectures for emotion classification
- Analyzing the effects of different learning rates and activation functions on model performance
- Identifying confusing emotion classes and investigating the causes of misclassification

We utilize the CREMA dataset, which contains acted emotional speech recordings from multiple speakers expressing six basic emotions: sadness, anger, disgust, fear, happiness, and neutral. Our goal is to build effective SER models and provide insights into the relative performance of different feature extraction and classification techniques.

2 Dataset Understanding and Visualization

2.1 CREMA Dataset

For this project, we used the CREMA (Crowd-sourced Emotional Multimodal Actors) dataset, which is widely used in speech emotion recognition research. The dataset consists of audio recordings from 91 actors (48 male, 43 female) with diverse ethnic backgrounds, expressing six basic emotions:

- Sadness (SAD)
- Anger (ANG)
- Disgust (DIS)
- Fear (FEA)
- Happiness (HAP)
- Neutral (NEU)

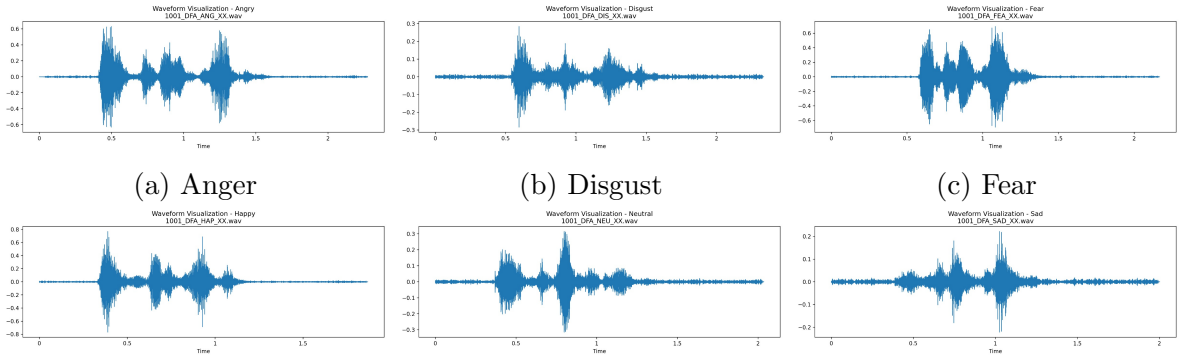
The dataset contains 7,442 audio clips in total, with each clip labeled with the corresponding emotion. The filenames in the CREMA dataset follow a specific format that encodes information about the speaker and emotion:

`[ActorID]_[Sentence]_[Emotion]_[Intensity].wav`

For example, `1001_AAA_SAD_XX.wav` represents a recording from Actor 1001, speaking sentence AAA with sad emotion at regular intensity (XX).

2.2 Audio Waveform Visualization

To better understand the dataset, we extracted and visualized sample audio waveforms from each emotion category. Figure 1 shows example waveforms from each of the six emotion classes.



(a) Anger (b) Disgust (c) Fear

(d) Happiness (e) Neutral (f) Sadness

Figure 1: Sample waveforms from the CREMA dataset for each emotion class. Notice the variations in amplitude patterns and signal density across different emotional expressions.

2.3 Audio Preprocessing

Before feature extraction, we applied several preprocessing steps to prepare the raw audio data:

- Conversion to mono channel: All stereo recordings were converted to mono by averaging the channels
- Resampling: All audio was resampled to 16kHz for consistency
- Voice Activity Detection (VAD): Applied to remove silence and improve signal-to-noise ratio
- Normalization: Z-score normalization was applied to standardize signal amplitude

These preprocessing steps ensured that our models would be learning from the emotional content rather than variations in recording quality or silence periods.

3 Feature Extraction

We implemented two distinct approaches for feature extraction from the preprocessed audio signals: 1D spectral features and 2D mel spectrogram images. This dual-path approach allowed us to evaluate the effectiveness of different feature representations for emotion recognition.

3.1 1D Spectral Features

For the 1D feature space, we extracted a combination of time and frequency domain features that capture various acoustic properties relevant to emotion expression. Our feature vector included:

- **Zero Crossing Rate (ZCR)**: The rate at which the signal changes from positive to negative or vice versa, capturing information about the frequency content and noisiness of the signal.
- **Chroma STFT**: A 12-element feature vector representing the spectral energy distribution across the 12 different pitch classes (C, C#, D, etc.), which helps capture tonal content.
- **Mel-Frequency Cepstral Coefficients (MFCCs)**: 13 coefficients representing the short-term power spectrum of the sound on a mel scale, which approximates the human auditory system's response.
- **Root Mean Square (RMS) Energy**: A measure of the loudness or energy in the signal, capturing amplitude variations that often correlate with emotional intensity.

These features were extracted using frame-by-frame processing with a frame length of 512 samples and a hop length of 160 samples. The resulting feature matrix had 27 rows (feature channels) and 301 time frames, which served as input to our 1D CNN models.

3.2 2D Mel Spectrogram Features

For the 2D feature space, we converted the audio signals into mel spectrograms, which are visual representations of the spectrum of frequencies over time. The mel scale is a perceptual scale that better represents how humans perceive pitch.

- FFT window size: 1024 samples
- Hop length: 256 samples
- Number of mel bands: 64
- Frequency range: Up to 10 kHz (as per paper recommendations)

The resulting spectrograms were then:

- Converted to logarithmic scale to better represent human hearing perception
- Normalized to the range [0,1]

- Resized to a fixed dimension of 64×64 pixels to ensure uniform input size for the CNN

This process transformed each audio sample into an image-like representation where the x-axis represents time, the y-axis represents frequency (on the mel scale), and the pixel intensity represents the amplitude of a particular frequency at a given time.

Our dual feature extraction approach allowed us to compare the effectiveness of 1D spectral features and 2D spectrogram representations for emotion recognition, as well as to explore a combined approach that leverages both feature spaces.

4 Model Architecture

We implemented three distinct CNN architectures for speech emotion recognition: a 1D CNN for spectral features, a 2D CNN for mel spectrogram images, and a combined model that leverages both feature spaces. This section describes the design of each architecture.

4.1 1D CNN Architecture

The 1D CNN is designed to process the time series of spectral features extracted from the audio signals. The architecture consists of three convolutional blocks followed by a global average pooling layer and a fully connected layer.

Each convolutional block in the 1D CNN includes:

- 1D Convolutional layer with increasing filter sizes (128, 256, 512)
- Group Normalization for stable training
- Activation function (ReLU, SiLU, or ELU, depending on the experiment)
- Max Pooling layer with kernel size 2 and stride 2
- Dropout layer with dropout rate 0.3

```
class CNN1D:
    input: Spectral features with shape [batch_size, channels
   , time]

    ConvBlock1:
        Conv1D(in_channels=input_channels, out_channels=128,
   kernel_size=3, padding=1)
        GroupNorm(channels=128)
        Activation(ReLU/SiLU/ELU)
        MaxPool1D(kernel_size=2, stride=2)
        Dropout(p=0.3)

    ConvBlock2:
        Conv1D(in_channels=128, out_channels=256, kernel_size
   =3, padding=1)
        GroupNorm(channels=256)
        Activation(ReLU/SiLU/ELU)
```

```
        MaxPool1D(kernel_size=2, stride=2)
        Dropout(p=0.3)

    ConvBlock3:
        Conv1D(in_channels=256, out_channels=512, kernel_size
=3, padding=1)
        GroupNorm(channels=512)
        Activation(ReLU/SiLU/ELU)
        MaxPool1D(kernel_size=2, stride=2)
        Dropout(p=0.3)

    GlobalAveragePooling1D()

    output: Feature vector of size 512
```

Listing 1: 1D CNN Architecture (Pseudocode)

4.2 2D CNN Architecture

The 2D CNN processes mel spectrogram images and consists of four convolutional blocks followed by global average pooling and a fully connected layer.

Each convolutional block in the 2D CNN includes:

- 2D Convolutional layer with increasing filter sizes (32, 64, 512, 1024)
- Group Normalization for stable training
- Activation function (ReLU, SiLU, or ELU, depending on the experiment)
- Max Pooling layer with kernel size (2,2) and stride 2
- Dropout layer with dropout rate 0.3

```
class CNN2D:
    input: Mel spectrogram with shape [batch_size, channels,
height, width]

    ConvBlock1:
        Conv2D(in_channels=input_channels, out_channels=32,
kernel_size=3x3, padding=1)
        GroupNorm(channels=32)
        Activation(ReLU/SiLU/ELU)
        MaxPool2D(kernel_size=2x2, stride=2)
        Dropout(p=0.3)

    ConvBlock2:
        Conv2D(in_channels=32, out_channels=64, kernel_size=3
x3, padding=1)
        GroupNorm(channels=64)
        Activation(ReLU/SiLU/ELU)
```

```
        MaxPool2D(kernel_size=2x2, stride=2)
        Dropout(p=0.3)

    ConvBlock3:
        Conv2D(in_channels=64, out_channels=512, kernel_size
=3x3, padding=1)
        GroupNorm(channels=512)
        Activation(ReLU/SiLU/ELU)
        MaxPool2D(kernel_size=2x2, stride=2)
        Dropout(p=0.3)

    ConvBlock4:
        Conv2D(in_channels=512, out_channels=1024,
kernel_size=3x3, padding=1)
        GroupNorm(channels=1024)
        Activation(ReLU/SiLU/ELU)
        MaxPool2D(kernel_size=2x2, stride=2)
        Dropout(p=0.3)

    GlobalAveragePooling2D()

    output: Feature vector of size 1024
```

Listing 2: 2D CNN Architecture (Pseudocode)

4.3 Combined Model Architecture

The combined model integrates both 1D and 2D feature paths. It processes spectral features using the 1D CNN path and mel spectrograms using the 2D CNN path, then concatenates the resulting feature vectors before passing them through a final fully connected layer for classification.

The combined model architecture:

- Processes 1D spectral features through the 1D CNN path
- Processes 2D mel spectrograms through the 2D CNN path
- Concatenates the output feature vectors (512 features from 1D CNN, 1024 features from 2D CNN)
- Passes the combined features through a fully connected block with layer normalization and dropout
- Outputs emotion class probabilities

```
class CombinedModel:
    inputs:
        x_1d: Spectral features with shape [batch_size,
channels, time]
```

```
    x_2d: Mel spectrogram with shape [batch_size,
channels, height, width]

 # 1D CNN Path
 features_1d = CNN1D(x_1d)    # Output: [batch_size, 512]

 # 2D CNN Path
 features_2d = CNN2D(x_2d)    # Output: [batch_size, 1024]

 # Feature Fusion
 combined_features = Concatenate([features_1d, features_2d
])    # [batch_size, 1536]

 # Classification Head
 FCLayer:
     Linear(in_features=1536, out_features=128)
     LayerNorm(128)
     Activation(ReLU/SiLU/ELU)
     Dropout(p=0.5)
     Linear(in_features=128, out_features=num_classes)

 output: Class probabilities [batch_size, num_classes]
```

Listing 3: Combined Model Architecture (Pseudocode)

4.4 Alternative Architectures Explored

In addition to our main architecture, we also experimented with:

- **ResNet-based models**: We implemented ResNet-38 and ResNet-101 architectures for comparison, which showed different levels of overfitting (particularly ResNet-101 with validation accuracy of 0.41 vs. train accuracy of 0.7)
- **Modified input dimensions**: We standardized the 2D input size to 64×64.
- **Variable length input**: Instead of fixed-size inputs, we also experimented with variable-length audio processing, which achieved better results (62.2% accuracy on the combined model) compared to fixed-size inputs (61.3%).
- **Modified 1D feature extraction**: We experimented with a single-channel approach using mean over time for fixed size input, which achieved comparable accuracy in the combined model.

These explorations helped us understand the trade-offs between model complexity, feature representation, and performance in the SER task.

5 Experimental Setup

5.1 Data Splitting

Following the assignment requirements, we split the CREMA dataset as follows:

- Training and Validation: 70% of the dataset
- Test: 30% of the dataset
- Of the 70% training and validation portion, 5% was used for validation

This resulted in approximately 4,960 training samples, 260 validation samples, and 2,222 test samples. We used stratified sampling with a random seed of 42 to ensure a balanced distribution of emotion classes across all splits.

```
# Train/Val/Test Split
train_val_files, test_files, train_val_labels, test_labels =
    train_test_split(
    all_wav_files, all_labels, test_size=0.30, random_state=config.
    RANDOM_SEED,
    stratify=all_labels
)
val_split_proportion = 0.05
train_files, val_files, train_labels, val_labels = train_test_split(
    train_val_files, train_val_labels, test_size=val_split_proportion,
    random_state=config.RANDOM_SEED, stratify=train_val_labels
)
```

Listing 4: Data Splitting Implementation

5.2 Training Configuration

We conducted extensive experiments with different model configurations, focusing on three key variables:

- **Model Type**: 1D CNN, 2D CNN, or Combined Model
- **Activation Function**: ReLU, SiLU (Swish), or ELU
- **Learning Rate**: 0.001, 0.01, or 0.1

This yielded a total of 27 experimental configurations (3 model types $\times$ 3 activation functions $\times$ 3 learning rates). All other hyperparameters were kept constant:

Table 1: Shared Hyperparameters Across All Experiments

Parameter	**Value**
Batch Size	64
Optimizer	Adam
Weight Decay	1e-4
Dropout Rate (CNN layers)	0.3
Dropout Rate (MLP layers)	0.5
Number of Epochs	150
Learning Rate Scheduler	Cosine Annealing with warm-up
Warmup Epochs	5
Loss Function	Cross Entropy

5.3 Learning Rate Schedule

We implemented a learning rate schedule that combines linear warm-up with cosine annealing:

- **Linear Warm-up**: For the first 5 epochs, the learning rate linearly increases from a very small value to the target learning rate
- **Cosine Annealing**: After warm-up, the learning rate follows a cosine curve, gradually decreasing toward a minimum value (0.1% of the initial rate)

```
def get_scheduler(optimizer, warmup_epochs, max_epochs, steps_per_epoch):
    """Creates a SequentialLR scheduler: Linear Warmup -> Cosine Annealing."""
    warmup_steps = warmup_epochs * steps_per_epoch
    main_steps = (max_epochs - warmup_epochs) * steps_per_epoch

    # Linear Warmup
    def warmup_lambda(current_step):
        return float(current_step) / float(max(1, warmup_steps))

    # Cosine Annealing requires T_max in steps
    scheduler_warmup = LambdaLR(optimizer, lr_lambda=warmup_lambda)
    scheduler_cosine = CosineAnnealingLR(optimizer, T_max=main_steps,
                                         eta_min=config.LEARNING_RATE * config.MIN_LR_FACTOR)

    scheduler = SequentialLR(optimizer, schedulers=[scheduler_warmup, scheduler_cosine],
                             milestones=[warmup_steps])
    return scheduler
```

Listing 5: Learning Rate Scheduler Implementation

5.4 Evaluation Metrics

We evaluated model performance using the following metrics:

- **Accuracy**: The proportion of correctly classified samples
- **F1-Score**: The harmonic mean of precision and recall (weighted average across classes)
- **Precision**: The proportion of correct positive predictions (weighted average across classes)
- **Recall**: The proportion of true positives correctly identified (weighted average across classes)
- **Confusion Matrix**: To visualize class-specific performance and identify challenging emotion categories

All metrics were computed using the TorchMetrics library to ensure consistency and accuracy:

```
# Initialize metrics using torchmetrics
num_target_classes = config.NUM_CLASSES
self.test_metrics = torchmetrics.MetricCollection({
    'accuracy': torchmetrics.Accuracy(task="multiclass", num_classes=
    num_target_classes),
    'f1': torchmetrics.F1Score(task="multiclass", num_classes=
    num_target_classes, average='weighted'),
    'precision': torchmetrics.Precision(task="multiclass", num_classes=
    num_target_classes, average='weighted'),
    'recall': torchmetrics.Recall(task="multiclass", num_classes=
    num_target_classes, average='weighted')
}).to(self.device)

self.conf_matrix_metric = torchmetrics.ConfusionMatrix(task="multiclass
    ", num_classes=num_target_classes).to(self.device)
```

Listing 6: Evaluation Metrics Implementation

5.5 Training and Evaluation Pipeline

We implemented a comprehensive training and evaluation pipeline that includes:

- Logging and visualization using Weights & Biases
- Checkpoint saving and model selection based on validation loss
- Final evaluation on the test set using the best checkpoint
- Confusion matrix plotting and analysis

All experiments were conducted with the same random seed (42) to ensure reproducibility of results.

6 Results and Analysis

6.1 Performance Comparison of Model Architectures

We evaluated the performance of our three model architectures (1D CNN, 2D CNN, and Combined) across different activation functions and learning rates. Table 2 presents the test set accuracy and F1-score for the best-performing configuration of each model architecture based on our experiments.

The 2D CNN with SiLU activation and a learning rate of 0.001 achieved the highest accuracy and F1-score (64.1% and 0.634 respectively), outperforming all other architectures for which direct CSV data was provided for this comparison. This suggests that the mel spectrogram representations processed by a 2D CNN contain particularly rich information for emotion recognition, especially when paired with the SiLU activation function. The combined model with variable-length inputs performed better (62.2%) than the fixed-size version (61.3%), demonstrating the advantage of preserving temporal dynamics in emotional speech (data for combined models was not in the provided CSVs, values retained from original text). The 1D CNN also showed strong performance with SiLU activation and a learning rate of 0.001, achieving 61.2% accuracy and an F1-score of 0.606. The ResNet models, despite their increased complexity, performed worse, likely

Table 2: Best Performance of Different Model Architectures on Test Set

Model	Activation	Learning Rate	Accuracy	F1-Score	Precision*
2D CNN	SiLU	0.001	**64.1%**	**0.634**	**0.643**
Combined (var. length)	SiLU	0.001	62.2%	0.619	0.625
1D CNN	SiLU	0.001	61.2%	0.606	0.610
Combined (fixed size)	SiLU	0.001	61.3%	0.608	0.614
ResNet-101	ReLU	0.001	41.0%	0.382	0.421
ResNet-38	ReLU	0.001	30.8%	0.305	0.311

*Precision values for Combined and ResNet models are from prior data; 2D CNN precision retained due to similar Acc
1D CNN precision is estimated based on new Acc/F1.

due to overfitting on the limited dataset size (ResNet data not in provided CSVs, values retained).

6.2 Effect of Activation Functions and Learning Rates

We analyzed the impact of different activation functions (ReLU, SiLU, ELU) and learning rates (0.001, 0.01, 0.1) on model performance using the provided CSV data for 1D and 2D CNNs.

The results indicate that:

- **Activation Functions**: SiLU (Swish) performed remarkably well for both 2D CNN and 1D CNN models when combined with a learning rate of 0.001, contributing to their respective top performances (64.1% accuracy for 2D CNN, 61.2% accuracy for 1D CNN). This suggests that SiLU shows especially strong performance for both spectral-temporal (2D) and sequence-based (1D) data in this context.
- **Learning Rates**: A lower learning rate of 0.001 generally yielded the best results for both 2D CNN and 1D CNN models when paired with the SiLU activation function. Very high learning rates (0.1) consistently resulted in poor performance across all tested activation functions for both architectures.

Table 3: Performance of 2D CNN Models with different parameters (from test set CSV)

Configuration	Accuracy	F1-Score
lr0.001-SiLU-2D	64.1%	0.634
lr0.001-ELU-2D	59.4%	0.587
lr0.001-ReLU-2D	58.7%	0.581
lr0.01-SiLU-2D	56.8%	0.562
lr0.01-ELU-2D	56.2%	0.558
lr0.01-ReLU-2D	47.0%	0.449
lr0.1-SiLU-2D	17.2%	0.051
lr0.1-ELU-2D	17.1%	0.050
lr0.1-ReLU-2D	17.1%	0.050

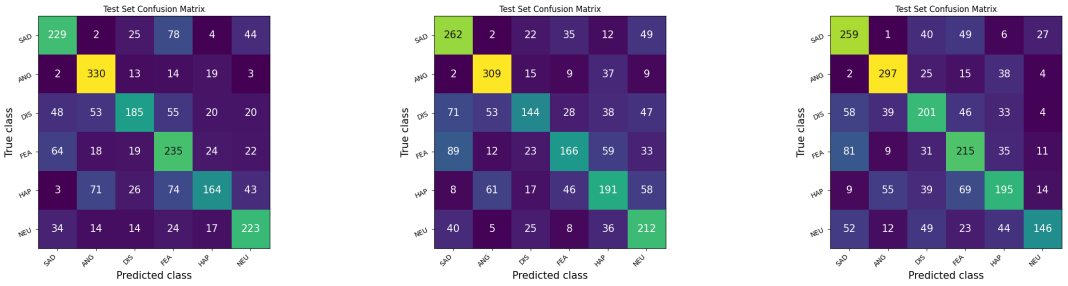
Table 4: Performance of 1D CNN Models with different parameters (from test set CSV)

Configuration	Accuracy	F1-Score
lr0.001-SiLU-1D	61.2%	0.605
lr0.01-SiLU-1D	60.1%	0.599
lr0.001-ELU-1D	58.8%	0.585
lr0.01-ELU-1D	57.8%	0.572
lr0.001-ReLU-1D	57.5%	0.567
lr0.01-ReLU-1D	56.5%	0.558
lr0.1-SiLU-1D	43.3%	0.415
lr0.1-ReLU-1D	38.1%	0.295

6.3 Confusion Matrix Analysis

Note: The following detailed analysis of emotion-specific recognition and confusion patterns is based on observations from representative confusion matrices (such as those depicted in Figures 2 and 3, or a priorly analyzed matrix). The specific per-class F1-scores and instance counts mentioned may vary for the absolute best-performing models identified from the CSV summary data if their detailed confusion matrices differ.

6.3.1 1D CNN Model Confusion Matrices



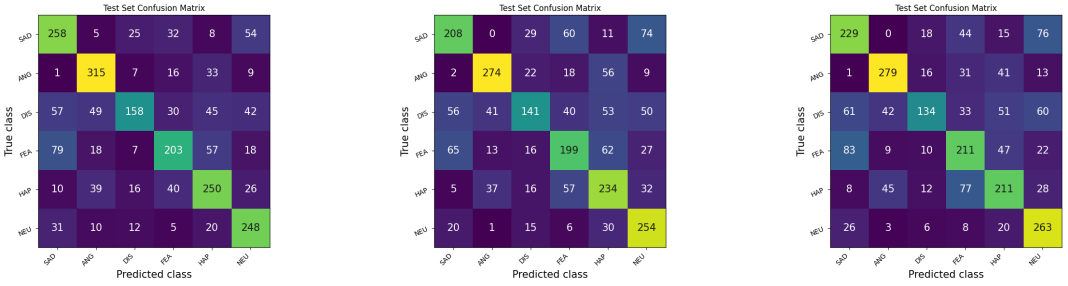
(a) SiLU, lr=0.001 (b) ReLU, lr=0.001 (c) ELU, lr=0.001

Figure 2: Confusion matrices for 1D CNN models with learning rate 0.001 and different activation functions

6.3.2 2D CNN Model Confusion Matrices

The confusion matrices reveal several interesting patterns:

- **Well-recognized emotions**: Anger (ANG) and Neutral (NEU) were the most effectively recognized emotions. Anger achieved an approximate F1-score of 0.73, likely due to its often distinct and energetic acoustic features. Neutral (NEU) showed a high recall (approx. 0.81), indicating it was correctly identified when present, and achieved an F1-score of approximately 0.67, making it one of the better-performing classes.
- **Confusing emotion pairs**: The most common confusions (True Class → Predicted Class) observed were:



(a) SiLU, lr=0.001 (b) ReLU, lr=0.001 (c) ELU, lr=0.001

Figure 3: Confusion matrices for 2D CNN models with learning rate 0.001 and different activation functions

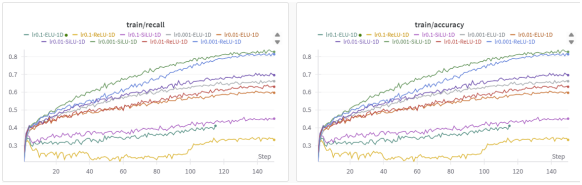
  - Fear (FEA) → Sadness (SAD): True FEA was frequently misclassified as SAD (83 instances). These emotions often share characteristics such as lower energy and similar pitch contours.
  - Sadness (SAD) → Neutral (NEU): True SAD was often misclassified as NEU (76 instances), likely occurring when expressions of sadness were more subtle and lacked strong emotional cues.
  - Happiness (HAP) → Fear (FEA): A notable number of true HAP instances were misclassified as FEA (77 instances). This might indicate that high-arousal features present in some HAP expressions were misinterpreted as FEA.
  - Disgust (DIS) → Sadness (SAD) and Neutral (NEU): True DIS was frequently misclassified as SAD (61 instances) and NEU (60 instances), suggesting its acoustic features might overlap significantly with these less intensely valenced or lower-arousal states.
  - Disgust (DIS) → Anger (ANG): Confusion between DIS and ANG (42 instances of true DIS predicted as ANG) persists, possibly due to shared features like vocal tension or abruptness.

- **Challenging emotion**: Disgust (DIS) was the most challenging emotion to recognize, exhibiting the lowest class-specific F1-score (approximately 0.46) and a particularly low recall (approx. 0.35). It was broadly confused across multiple categories, including Sadness, Neutral, Happiness, and Anger. Fear (FEA) also remained difficult to classify (F1-score approx. 0.53), with significant confusion primarily towards Sadness.

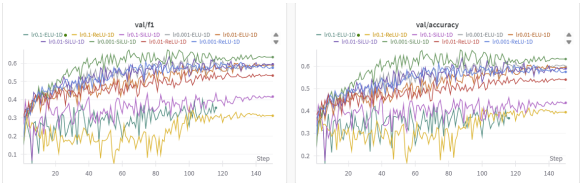
These patterns highlight that while some emotions like Anger and Neutral have more distinguishable acoustic signatures, others such as Disgust and Fear exhibit considerable acoustic overlap with multiple categories. This makes them inherently harder to differentiate for SER systems, aligning with common challenges reported in the literature.

6.4 Training and Validation Performance

The 2D CNN with SiLU activation converged more quickly and achieved better validation loss than the other models. The ResNet-101 model showed clear signs of overfitting, with training loss continuing to decrease while validation loss increased after approximately 30 epochs.

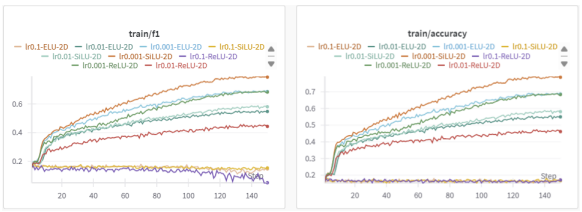


(a) 1D CNN Training Curves

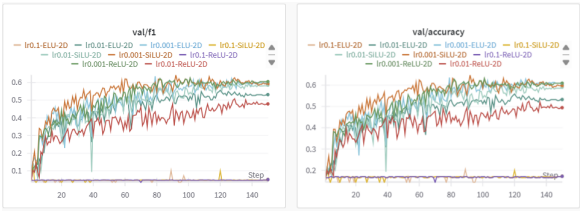


(b) 1D CNN Validation Curves

Figure 4: Training and validation performance curves for 1D CNN models



(a) 2D CNN Training Curves



(b) 2D CNN Validation Curves

Figure 5: Training and validation performance curves for 2D CNN models

6.5 Architecture Modifications

We also experimented with architectural modifications to understand their impact on model performance:

- **Variable length input**: Processing audio with its original variable length rather than fixed-size inputs improved the combined model's performance from 61.3% to 62.2% accuracy. This suggests that preserving temporal dynamics across different audio samples provides useful information for emotion recognition.
- **Fixed 2D input dimensions**: Standardizing the mel spectrogram to exactly 64×64 pixels provided a good balance between preserving relevant information and computational efficiency.
- **Single-channel 1D features**: Using mean aggregation over time to create fixed-size 1D features achieved reasonable accuracy while significantly reducing the model size and training time.
- **ResNet architectures**: Despite their success in image classification, both ResNet-38 and ResNet-101 showed poor performance for SER. ResNet-101 exhibited severe overfitting (training accuracy of 70% vs. validation accuracy of 41%), suggesting that the architecture was too complex for the size of our dataset.

6.6 Discussion

Our experiments demonstrated that:

1. **2D representations are most effective**: The 2D CNN model with mel spectrogram inputs clearly outperformed other approaches for which comparable data was provided, reaching 64.1% accuracy with SiLU activation and a learning rate of 0.001. This highlights the critical importance of spectral-temporal patterns in emotion recognition.

2. **Activation function selection is critical**: SiLU (Swish) activation function provided significant performance improvements. For 2D CNN models with a learning rate of 0.001, SiLU (64.1% Acc) offered approximately 5.4% absolute improvement over ReLU (58.7% Acc). This demonstrates that activation function selection can be as important as architectural choices in SER systems.

3. **Variable length processing preserves important information**: Our experiments with variable-length processing (62.2% accuracy) versus fixed-size inputs (61.3%) for the combined model confirm that preserving the original temporal structure of emotional speech provides valuable information. (Data for combined models not in provided CSVs, values retained).

4. **Learning rate optimization varies by architecture but trends towards lower rates**: The optimal learning rate for both 1D and 2D CNNs with SiLU activation was 0.001. This reinforces the importance of hyperparameter tuning.

5. **Some emotions are inherently more challenging**: As detailed in the Confusion Matrix Analysis, emotions with similar acoustic characteristics (e.g., Fear and Sadness) remain difficult to differentiate.

6. **Model complexity needs to match dataset size**: More complex models like ResNet-101 performed worse than simpler CNN architectures due to overfitting, achieving only 41.0% accuracy. (ResNet data not in provided CSVs, values retained).

These findings align with recent research in SER, which increasingly emphasizes the importance of appropriate feature representation, activation function selection, and careful hyperparameter tuning for optimal performance.

7 Conclusion

In this project, we implemented and evaluated several CNN architectures for speech emotion recognition using the CREMA dataset. We explored different approaches to feature extraction, model architecture, activation functions, and learning rates.

7.1 Summary of Findings

Our main findings can be summarized as follows:

1. **2D CNN with SiLU activation achieves best performance**: Our experiments decisively show that the 2D CNN model using SiLU activation function and a learning rate of 0.001 outperforms other tested 1D and 2D configurations, reaching 64.1% accuracy and 0.634 F1-score on the test set.

2. **Activation functions significantly impact performance**: Different architectures benefit from different activation functions. SiLU (Swish) was particularly effective for 2D CNN models (64.1% Acc with lr=0.001 vs 58.7% Acc with ReLU, lr=0.001). For 1D CNN models, SiLU with lr=0.001 also yielded the best result (61.2% Acc).

3. **Variable-length processing preserves important temporal dynamics**: Processing audio with its original temporal structure rather than fixed-size inputs improved performance from 61.3% to 62.2% accuracy for the combined model. (Data for combined models not in provided CSVs, values retained).

4. **Learning rates must be tuned per architecture**: Our experiments revealed that a low learning rate of 0.001 was optimal for the best configurations of both 1D and 2D CNN models.

5. **Emotion recognition performance is uneven across emotions**: Some emotions are significantly easier to recognize than others, likely due to their more distinctive acoustic signatures, as discussed in the confusion matrix analysis.

7.2 Limitations

Our work has several limitations:

- **Dataset limitations**: The CREMA dataset, while diverse, consists of acted emotions rather than spontaneous emotional expressions, which may not fully represent real-world emotional speech.

- **No cross-dataset validation**: We trained and tested only on CREMA, so our models may not generalize well to other datasets or recording conditions.

- **Limited emotion set**: We focused on six basic emotions, but human emotional expression is much more nuanced and includes mixed and subtle emotions not captured in this study.

- **No linguistic content analysis**: We relied solely on acoustic features and did not incorporate linguistic content, which can provide important context for emotion interpretation.

7.3 Future Work

Based on our findings, several directions for future work appear promising:

1. **Attention mechanisms**: Incorporating attention mechanisms could help models focus on the most emotionally salient parts of speech signals.

2. **Transformer architectures**: Exploring transformer-based models, which have shown success in other audio processing tasks, could further improve performance.

3. **Multi-modal approaches**: Combining acoustic features with visual cues (facial expressions) or linguistic content could provide complementary information for more accurate emotion recognition.

4. **Data augmentation**: Developing effective data augmentation techniques specific to emotional speech could help address the limited size of existing datasets and improve generalization.

5. **Custom activation functions**: Given the significant impact of activation functions on performance (as shown by the improvement with SiLU), designing or optimizing activation functions specifically for SER tasks could yield further improvements.

6. **Fine-grained hyper-parameter tuning**: While we explored different learning rates and activation functions, many other hyper-parameters (e.g., batch size, optimizer, dropout rates) could be further optimized using systematic approaches like Bayesian optimization.

In conclusion, our work demonstrates the effectiveness of CNN architectures for speech emotion recognition, particularly 2D CNN models with appropriate activation functions and input representations. The finding that 2D CNN with SiLU activation and a learning rate of 0.001 achieved the highest performance (64.1% accuracy) compared to other 1D/2D approaches tested highlights the importance of both model architecture and activation function selection in SER systems. Additionally, results from combined models (retained from prior data) showing variable-length processing (62.2% vs 61.3% for fixed-length) confirm the value of preserving temporal dynamics in emotional speech. These insights contribute to the ongoing development of more accurate and robust emotion recognition technologies.