*ECAI 2024*
*U. Endriss et al. (Eds.)*
*© 2024 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA240774*

# Benchmarking Adversarial Robustness in Speech Emotion Recognition: Insights into Low-Resource Romanian and German Languages

**Sebastian-Vasile Echim**[a], **Răzvan-Alexandru Smădu**[a] **and Dumitru-Clementin Cercel** [a,*]

[a]Faculty of Automatic Control and Computers, National University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania

**Abstract.** Therapy, interviews, and emergency services assisted by artificial intelligence (AI) are applications where speech emotion recognition (SER) plays an essential role, for which performance and robustness are subject to improvement. Deep learning approaches have proven effective in SER; nevertheless, they can underperform when exposed to adversarial attacks. In this paper, we explore and enhance architectures, such as convolutional neural networks with long short-term memory (CNN-LSTM), AlexNet, VGG16, Convolutional Vision Transformer (CvT), Vision Transformer (ViT), and LeViT, by finding the suitable setup for SER models regarding speech processing, network hyperparameters, spectrogram augmentations, and adversarial examples. We apply our methodology to Romanian and German SER datasets and achieve state-of-the-art results, with 89.81% validation weighted accuracy and 98.09% average weighted accuracy on the trained models. Our highly robust models reach complete adversarial defense and up to 5.56% weighted accuracy improvement when attacked. We also show how adversarial attacks influence model behavior in SER through explainable AI techniques.

## 1 Introduction

Speech represents the most effective way for humans to communicate and express themselves. As technology evolved, the necessity of interacting with computers grew significantly. Various interfaces have been developed for human-machine interaction, considering tangible (e.g., keyboards and mice) and non-tangible (e.g., gesture and vision-based interfaces) [29]. Intensively explored over the past decades, human-machine interaction via speech is highly regarded as one of the most efficient interaction methods [14]. Such systems include automatic speech recognition [37], which recently has seen significant improvements [40]. Automatic speech recognition systems extract information from spoken utterances and convert them into sequences of words. Speech emotion recognition (SER) has been used to extract the speaker's emotional state to make the human-machine interaction more natural [38]. Automatic SER systems cover applications such as tools for therapists providing aid for diagnosing patients, AI-assisted emergency systems [43, 45], content-streaming, and interview platforms. The application of SER is also generally perceived as a means by which the user's voice adapts a system's behavior [14].

Speech processing has been explored for 1-dimensional audio waveforms [52] using deep neural network architectures such as 1D convolutional neural networks (CNN) or 1D CNNs together with recurrent neural networks (RNNs) [39], such as long short-term memory (LSTM) [17]. However, to take full advantage of the information provided, a 2D feature representation of the speech signal passed to 2D networks showed better performance [1, 52]. Speech processing can be seen from a computer vision standpoint by transforming the speech waveform to a visual representation using speech features denoting the acoustic strength of the signal over the time and frequency domains. They are obtained by computing the short-time Fourier transform that determines discrete Fourier transforms over a specified number of overlapping windows [23]. The most common types of 2D speech representations are linear, log, and mel-scaled spectrograms [31, 41].

Szegedy et al. [44] discovered that deep neural networks (DNNs) are susceptible to small, humanly imperceptible perturbations in the input data, causing wrong class assignations. This negative effect of small data changes on classifiers is known as an adversarial attack. Despite many adversarial defense algorithms in the literature [49], they are proven to have only local applicability. Therefore, the resistance of DNNs to malicious attacks is still an open subject. One such attack is based on adversarial examples introduced by Goodfellow et al. [15], augmentations generated by applying crafted perturbations to input samples. Adversarial attacks are split into three knowledge classes [34, 50]: white-box, grey-box, and black-box attacks. White-box attacks know the target of the attack and have access to the entire model. Black-box attacks are crafted using only the model's output or involving no model information. The grey-box attacks only access the target and generate examples by training a generative model [50]. Given the limited resources, these attacks are breaking the security of well-defended models [51].

Our paper introduces robust networks for SER in a Romanian corpus of emotional utterances, namely Emo-IIT [32]. Additionally, the German Berlin Database of Emotional Speech (Emo-DB) [7] is used in our base experiments to ensure task objectivity in our setup. Our investigations cover training on inputs represented in four different types of spectrograms: short-time Fourier transform (linear), mel-scaled (mel), constant-Q transform (CQT) [6], and mel-frequency cepstral coefficients (MFCC). We evaluate common architectures based on CNN and RNN [52], AlexNet [21], VGG16 [42], Vision Transformer (ViT) [13], Convolutional Vision Transformer

[*] Corresponding Author. Email: dumitru.cercel@upb.ro.

(CvT) [48], and faster inference Vision Transformer (LeViT) [16]. We assess the susceptibility of the obtained models to adversaries using a set of white-box adversarial algorithms, namely fast gradient sign method (FGSM) [15], FGSM with momentum (MI-FGSM) [12], basic iterative method (BIM) [22], projected gradient descent (PDG) [28], and expectation over transformation projected gradient descent (EOT+PGD) [3, 27], and two black-box adversarial algorithms, namely Pixle [36] and Square Attack [2]. The adversarial defense capability is also performed by training with the white-box algorithms and testing the white-box and black-box algorithms. Finally, we assess the $\epsilon$-sensitivity for the adversarial algorithms.

In this work, we bring the following main contributions:

- We thoroughly experiment with various spectrogram features used in diverse network architectures, also depicting the classification using t-SNE plots [46].
- We investigate the susceptibility of adversarial examples for the introduced networks through their adversarial testing performance metrics.
- We make a significant step into explaining the behavior of the DNNs on adversarial attacks and defenses.
- We obtain robust networks for white-box adversarial algorithms.
- We assess the black-box attack success rate for white-box defended models.

## 2 Related Work

### 2.1 Speech Processing

Badshah et al. [4] used linear spectrogram features for SER on the Emo-DB dataset, on which they trained a CNN architecture from scratch and fine-tuned AlexNet. Zhao et al. [52] introduced two local and global hierarchical architectures feature learning: a 1-dimensional CNN-LSTM model on raw audio clip features and a 2-dimensional CNN-LSTM on the spectrogram representation. They showed that introducing log-mel spectrograms in the CNN-LSTM setups for speech emotion recognition outperformed raw audio data by 15% validation accuracy in speaker-dependent experiments and 25% validation accuracy in speaker-independent experiments.

Muller et al. [33] performed spoof detection with constant-Q transform, log-scaled, and mel-scaled 513-dimensional spectrogram features created with the librosa library [30]. They evaluated twelve different models in fixed 4-second and variable input configurations. In addition, they experimented with models on raw waveforms, which outperformed the feature-based models due to the finer feature-extraction resolution for the raw inputs. They observed that mel-scaled spectrograms were underperforming compared to log-scaled and CQT spectrograms while replacing the mel spectrograms with CQT attained a 37% performance improvement. Chen et al. [9] introduced SpeechFormer in a setup consisting of linear and log-mel spectrograms and wav2vec [40] acoustic features. Their vanilla transformer-based [47] framework consisted of four sets of SpeechFormer blocks representing four stages interleaved by three merging blocks. The modeling process involved converting from frame to phoneme, word, and utterance. The test results on four datasets show state-of-the-art results with an improvement of 0.5-3.5% for weighted accuracy and 4-11% for unweighted accuracy.

In the Romanian language, Ungureanu et al. [45] introduced an emergency system architecture featuring an automatic speech recognition model using MFCC features extracted from 25 ms frames with a 10 ms stride as input. They involved a deep neural network consisting of 12 Time Delay Neural Network Factorization recurrent layers, which was trained on 394 hours and then evaluated on 62 hours of audio. While testing on four subsets, they achieved a 2.99 to 5.94 word error rate (WER). A SER model is defined in the emergency system using log-scaled spectrograms from 1-second audio segments created by splitting speech files from the Emo-IIT dataset [32] with a 10 ms stride. A dynamic range normalization within $[-90, -7]$ decibels is applied to the features, transforming them into 3-channel images. These resulting spectrograms were passed to a pre-trained VGG16 model. Speech emotion recognition results improved the baseline performance on weighted accuracy by 5-7%.

### 2.2 Adversarial Attack and Defense

White-box adversarial attacks have gained popularity since Goodfellow et al. [15] introduced a family of fast methods for generating adversarial examples, including the well-known FGSM. Their observations showed that malicious samples improved model generalization and feature regularization better than dropout. Kurakin et al. [22] showed that adversarial attacks are possible in the real world, on camera-taken images, by extending the fast method when introducing the BIM, attaining better attack success rate than FGSM on various perturbation values. Madry et al. [28] assessed the success of the PGD in obtaining robust networks. The results showed better attacks with PGD; however, when training using PGD adversarial examples on the MNIST dataset [25], the model was more resistant to PGD attacks but significantly decreased the accuracy on standard and FGSM test data. Ilyas et al. [18] introduced a set of black-box adversarial attacks to improve the success rate by a limited number of queries. The algorithms were based on three different settings: a query-efficient setup proving 2-3 times fewer queries for targeted adversarial attacks; a partial-information configuration, denoting the access to probabilities in top-k classes; and a label-only setup, representing an ordered list of predicted probabilities. The algorithms introduced showed successful attacks on real-world systems.

In the monolingual, multilingual, and cross-lingual speaker recognition settings, Liao et al. [26] employed FGSM and PGD attacks on ResNet-18, ViT, and CvT architectures, revealing the susceptibility of DNNs to attacks in speaker recognition tasks. They noticed that most models exhibited attack weaknesses, but the MFCC features were the most vulnerable to perturbations; however, no defense strategy, such as adversarial training, was applied to test the defense capability of their models.

## 3 Methodology

### 3.1 Dataset

In this paper, we rely on two speech emotion recognition datasets: the Romanian Emo-IIT dataset [32], which covers 522 audio files representing utterances denoting different mental states, and the German Emo-DB dataset [7], which covers 535 audio files. Both share the same classes (i.e., anger, boredom, disgust, fear, happiness, neutral state, and sadness).

**Emo-IIT.** Firstly introduced by Monica et al. [32], Emo-IIT is a Romanian resource of speech processing representing sentences of 20-22-year-old students, recorded at a 16KHz sample frequency. The total number of recordings was initially over 12,000. The validation was performed by 20 non-professional persons, resulting in two collections: one containing 2,994 files for which the emotion was recognized by more than 50% of people and another containing 2,502 files for 75% emotion recognition. However, the number of files made

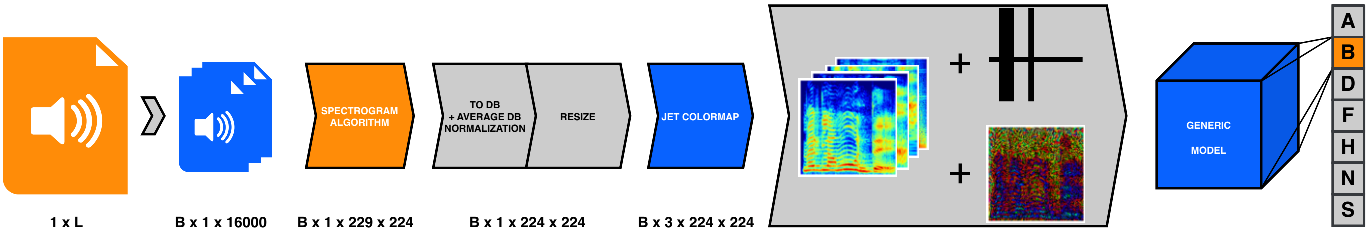


**Figure 1**: The speech processing pipeline considers an audio file of length $L$. $B$ is the resulting number of files trimmed to 1-second audio files after applying an audio cut with stride $N_{stride}$ in milliseconds (see Equation 1, $S_R = 16KHz$). After applying the spectrogram algorithm and converting it to a jet colormap, we use adversarial augmentation strategies and feed the resulting image into the neural network. The output of the network is the emotion identified in the speech.

available was 522, for only 3 out of 6 sentences; thus, we use this reduced dataset in our experiments.

**Emo-DB.** Recorded in 1997 and 1999 and presented by Burkhardt et al. [7], the Emo-DB is a collection of around 800 utterances of five female and five male voices, recorded in an anechoic chamber at the 48KHz sample rate, which was downsampled to 16KHz. The evaluation, done between 1997 and 2005, used sentiment and natural tone identification to reduce the database to 535 audio files.

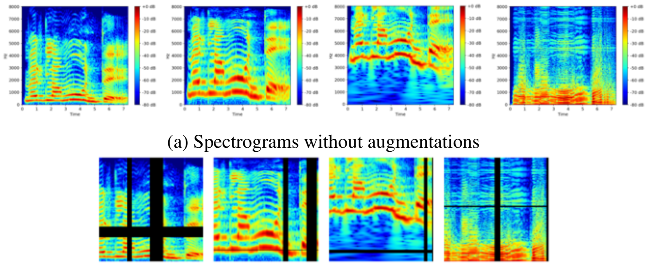
### 3.2 Speech Processing

We illustrate the processing pipeline in Figure 1. First, we split the datasets of 522 and 535 files, respectively, into 1-second samples. Similar to the work of Lech et al. [24], the resulting number of audio splits $B$ is determined by $N_{stride}$, as shown in Equation 1.

$$B = ceil\left(\frac{\left(\frac{max(L,S_R+1)}{S_R[Hz]} - 1[s]\right) * 1000}{N_{stride}[ms]}\right) \tag{1}$$

If the audio length $L$ exceeds the sample rate $S_R = 16,000$, the resulting 1-second file audio is padded with 0s to the right. Next, we pass the resulting files to four types of spectrogram algorithms using the librosa library [30]: linear, mel, CQT, and MFCC with 70 samples between successive frames, and a window size of 256, chosen such that the output is closer to the desired model input. The result is transformed in decibels and normalized based on the average minimum and maximum decibels over the entire dataset. Once the files are normalized, they are resized from $B \times 1 \times 229 \times 224$ to $B \times 1 \times 224 \times 224$, then passed to the jet colormap converter to determine the 3-channel images of shape $B \times 3 \times 224 \times 224$. Figure 2a depicts an example of the resulting spectrograms.

As a regularization technique, we perform an augmentation method similar to SpecAugment [35]. Therefore, our augmentation stretches the jet colormap image by a factor of 0.9 and applies two-time masks of 16 and a frequency mask of 16. Figure 2b shows an example for each type of spectrogram used.



(a) Spectrograms without augmentations

(b) Spectrograms with augmentations

**Figure 2**: Linear, mel, CQT, and MFCC spectrogram examples for the Emo-IIT file "B303fJa.wav".

### 3.3 Adversarial Algorithms

To assess and improve the robustness of our networks, we employ a set of white-box and black-box adversarial algorithms using Kim's adversarial toolbox, namely torchattacks [19]. Thus, we utilize five gradient-based white-box algorithms [5]: FGSM, MI-FGSM, BIM, PGD, and EOT+PGD. They feature several hyperparameters, including a perturbation factor $\epsilon$ for the weight of noise that is applied, which is subject to variations for the attack impact assessment of the algorithms. We denote $x'_0$ the initial adversarial example, $x'_t$ the adversarial example after t-steps, $L$ the loss function, $N_{steps}$ the number of algorithm steps, $\alpha$ the step size, and $\epsilon$ the maximum perturbation.

**FGSM.** A simple yet effective approach is to find adversarial examples that maximize the loss function. These examples are generated in a gradient ascend-like manner by adding a small perturbation proportional to the sign of the gradient of the loss function [15]:

$$x' = x + \epsilon \cdot \text{sgn}\left(\nabla_x L(f(x), y)\right) \tag{2}$$

The perturbation is subject to an $\ell_\infty$ constraint such that $||x' - x||_\infty < \epsilon$, resulting in adversarial examples $x'$ close to the input points $x$.

**PGD.** An extension to the fast gradient method is to apply an iterative process in which we use a small perturbation to the input $x$, every step [28]. Because the step-by-step process accumulates the perturbation factor, resulting in significant changes in the input space, a projection operator $\Pi_{\mathcal{B}(x,\epsilon)}$ is applied to the $\epsilon$-ball $\mathcal{B}(x, \epsilon)$. The adversarial sample generation process is formalized below:

$$\begin{aligned} x'_0 &= x + \mathcal{U}(-\epsilon, \epsilon) \\ x'_{t+1} &= \Pi_{\mathcal{B}(x,\epsilon)}\left\{x'_t + \alpha \cdot \text{sgn}\left(\nabla_{x'_t} L\left(f\left(x'_t\right), y\right)\right)\right\}, \end{aligned} \tag{3}$$

where $\mathcal{U}$ is the uniform distribution.

**MI-FGSM.** Because the iterative process may lead to poor local maxima, momentum-based iterative methods have been proposed to boost performance by accumulating gradients at each iteration, with a decay factor $\mu$ [12]. The recurrence is defined as follows:

$$\begin{aligned} g_{t+1} &= \mu \cdot g_t + \frac{\nabla_{x'_t} L\left(f\left(x'_t\right), y\right)}{\left\|\nabla_{x'_t} L\left(f\left(x'_t\right), y\right)\right\|_1}, \\ x'_{t+1} &= \Pi_{\mathcal{B}(x,\epsilon)}\left\{x'_t + \alpha \cdot \text{sgn}\left(g_{t+1}\right)\right\} \end{aligned} \tag{4}$$

where $\Pi_{\mathcal{B}(x,\epsilon)}$ is the projection to $\epsilon$-ball $\mathcal{B}(x, \epsilon)$ to keep the perturbation small enough while still fooling the network.

**BIM.** Another way to maintain small perturbations in the iterative methods is by clipping between $-\epsilon$ and $\epsilon$ [22]. Therefore, the iterative FGSM with $\epsilon$-neighborhood constraints is:

$$x'_{t+1} = \text{clip}_{(x,\epsilon)} \left\{ x'_t + \alpha \cdot \text{sgn}\left( \nabla_{x'_t} L\left(f\left(x'_t\right), y\right)\right)\right\}, \quad (5)$$

where $\text{clip}_{(x,\epsilon)}\{x'\} = \min\left(\max\left(x', x-\epsilon, 0\right), x+\epsilon, 255\right)$ assures the values are in the $\ell_\infty$ $\epsilon$-neighborhood of the original image and between 0 and 255.

**EOT+PGD.** Adversarial examples constructed with the previous techniques may lose their property when transformations are applied to the input [53]. For example, 2D/3D pose transformations such as translation, rotation, and scale applied to an adversarial image may yield the correct class. Therefore, to address this issue, we can introduce a distribution of such transformations for which the attack is invariant, and then we try to generate perturbations that produce valid adversarial examples based on expectation over transformations [53]:

$$x'_{t+1} = \Pi_{\mathcal{B}(x,\epsilon)} \left\{ x'_t + \alpha \cdot \text{sgn}\left( \mathbb{E}\left[ \nabla_{x'_t} L\left(f\left(x'_t\right), y\right)\right]\right)\right\}, \quad (6)$$

where $f$ is a randomized model generating a different output for each forward regardless of the identity of the input. The implementation in Zimmermann [53] uses $\frac{1}{m}\sum_i^m \nabla_{x'_t} L\left(f\left(x'_t\right), y\right)$ as an approximation of $\mathbb{E}\left[\nabla_{x'_t} L\left(f\left(x'_t\right), y\right)\right]$.

**Pixle and Square Attack.** Independent of the gradient information, Pixle [36] is based on randomly sampling parameterized patches of pixels and rearranging them in other positions. The Square Attack algorithm [2] provides $\ell_2$ and $\ell_\infty$-score-based attacks. Given a square size, the elements of patches are modified based on probability.

### 3.4 Models

We introduce six different architectures in our experiments with a diverse range of parameters and different input processing methods. The CNN-LSTM network [52] is trained on our chosen datasets from scratch. This network was set up in our experiments with feature extraction submodules composed of a 2D convolution, a batch normalization layer, an ELU activation function [10], and a 2D max pooling layer. Moreover, AlexNet [21], VGG16 [42], CvT [48], ViT [13], and LeViT [16] networks are pre-trained on ImageNet-1k [11] and used for fine-tuning.

## 4 Experimental Setup

### 4.1 Model Hyperparameters

For the CNN-LSTM model, as 3-channel images represent our spectrograms, we build the five feature extraction submodules with the input and output convolutional pairs of [(3, 32), (32, 64), (64, 128), (128, 128)], a kernel size of 2 and stride of 1 for the convolutional layers, a kernel size of 2 and a stride of 2 for the max-pooling layers, a LSTM input size of 169, and a LSTM hidden dimension of 256. The output of CNN-LSTM layers is passed through two fully connected layers from 32,768 to 1,000, then reduced to the number of classes (7), and a dropout of 0.4 is applied to the first layer. Since the other models are pre-trained on ImageNet-1k, we include a classification head represented by a fully connected layer from 1,000 to 7. The number of parameters for each model is 33.32M for CNN-LSTM, 61.11M for AlexNet, 138.36M for VGG16, 31.63M for CvT, 86.57M for ViT, and 18.90M LeViT.

### 4.2 Adversarial Hyperparameters

All adversarial algorithms are set up with $\epsilon = 32/255$. BIM, MI-FGSM, PGD, and EOT+PGD are defined with a step size $\alpha = 4/255$ and a number of steps $N_{steps} = 2$. MI-FGSM is also initialized with the gradient decay $\mu = 0.9$. EOT+PGD is configured with two EOT iterations. For the black-box attack algorithms, Pixle is set up with patch bounds of (20, 50) on both the X and Y axis, a maximum of one restart, and one iteration per restart. Square Attack is initialized with a maximum of 200 queries and one restart. For the experiments involving the variation of attack intensity, we choose in our tests $\epsilon \in \{8, 16, 32, 64, 128, 196\}/255$.

### 4.3 Evaluation and Training Hyperparameters

We quantitatively evaluate the performance of our models on metrics such as weighted precision, recall, validation accuracy, average accuracy (merged train and test classification accuracy, matching the baseline [45, 52] metrics), and confusion matrix. We compute the success rate for adversarial attacks, representing the percentage of accuracy reduction between testing with regular and adversarial examples. We also perform a qualitative analysis using t-SNE embedding plots [46] to visually explain the performance of the models and adversarial attacks along with Grad-CAM++ [8]. The optimizer is Adam [20], and the loss function is cross-entropy. We set the learning rate to $1e-4$ with a weight decay of $1e-3$ and train for 15 epochs. The batch size is 32, and the train/test split is 80%/20% random split.

## 5 Results

### 5.1 Preliminary Results

**Dataset Comparison.** Table 1 reveals the preliminary results on Emo-DB and Emo-IIT datasets. The overall performance on both datasets is similar. The best-performing model (i.e., CvT) achieves the highest accuracy on the Emo-IIT dataset when employing linear spectrogram inputs. In contrast, the same model architecture achieves the highest scores on the Emo-DB dataset when utilizing mel spectrogram inputs.

Despite having longer samples, training on the Emo-DB dataset does not yield better results when using the exact input representation. The total length is 1,487 seconds for the Emo-DB dataset, as well as 649 seconds for the Emo-IIT, which is less than half of Emo-DB. Moreover, considering the file number is similar between the datasets, we expect better performance for more speech data. CvT obtained the best validation weighted accuracy of 86.82% and an average weighted accuracy of 97.38% with mel spectrogram inputs on Emo-DB, as well as 88.12% validation accuracy and 97.75% average accuracy on Emo-IIT when employing linear spectrograms. Overall, performance is not different between the selected datasets, with a p-value $< 1\%$ using the paired t-test statistic.

**Model Comparison.** Regarding the input type, the best performance on Emo-DB is achieved using mel-scaled spectrograms with a 1.40% validation accuracy and 0.3% average accuracy over the second-performing spectrogram type, CQT. Meanwhile, Emo-IIT testing is better with linear spectrograms, with an added value of 0.6% for validation accuracy, compared to the CQT spectrogram and an improvement of 0.11% over linear spectrogram type. CQT consistently achieves the highest scores on both datasets, with validation accuracy of 79.86% and 82.16% on Emo-DB and Emo-IIT, respectively, and a mean of 95.30% and 96.56% on average accuracy.

**Table 1**: Baseline validation and average weighted accuracy results on datasets, spectrogram types, and deep neural network architectures. Higher values (↑) represent better performance, indicated in bold.

| Spectrogram | Model | Valid. Acc. (%) ↑ | | Avg. Acc. (%) ↑ | |
|---|---|---|---|---|---|
| | | Emo-DB | Emo-IIT | Emo-DB | Emo-IIT |
| LINEAR | CNN-LSTM | 78.52 | 77.24 | 95.76 | 95.75 |
| | AlexNet | 78.72 | 84.35 | 95.74 | 97.05 |
| | VGG16 | 79.85 | 81.09 | 96.11 | 96.49 |
| | CvT | **84.79** | **88.12** | **96.99** | **97.75** |
| | ViT | 75.22 | 77.41 | 90.55 | 94.88 |
| | LeViT | 71.65 | 81.88 | 93.26 | 96.60 |
| MEL | CNN-LSTM | 79.59 | 81.54 | 95.98 | 96.55 |
| | AlexNet | 78.32 | **87.44** | 95.71 | **97.64** |
| | VGG16 | 81.95 | 83.06 | 96.50 | 96.85 |
| | CvT | **86.82** | 82.18 | **97.38** | 96.57 |
| | ViT | 75.55 | 73.22 | 93.72 | 91.13 |
| | LeViT | 70.44 | 74.58 | 91.80 | 94.53 |
| CQT | CNN-LSTM | 79.65 | 80.97 | 95.96 | 96.44 |
| | AlexNet | 83.54 | 86.60 | 96.72 | 97.47 |
| | VGG16 | 81.17 | 83.23 | 96.26 | 96.86 |
| | CvT | **85.42** | **87.52** | **97.08** | **97.63** |
| | ViT | 76.83 | 73.24 | 93.89 | 94.94 |
| | LeViT | 72.56 | 81.41 | 91.88 | 96.02 |
| MFCC | CNN-LSTM | 70.57 | 76.07 | 94.17 | 95.51 |
| | AlexNet | 76.27 | **83.38** | 95.33 | **96.87** |
| | VGG16 | 77.45 | 80.17 | 95.52 | 96.31 |
| | CvT | **80.52** | 81.05 | **96.18** | 96.43 |
| | ViT | 68.85 | 70.38 | 85.93 | 94.06 |
| | LeViT | 73.34 | 75.36 | 92.65 | 95.15 |
| MEL + CNN-LSTM [52] | | 82.42 | - | 95.89 | - |
| LOG + VGG16 [45] | | - | - | 94.46 | 94.98 |

The comparison of different architectures shows better overall performance with AlexNet and CvT. For the linear and CQT spectrogram types, CvT is the best architecture, with top scores. For the mel-scaled and MFCC types, on Emo-IIT validation accuracy, AlexNet has an added value of 4.38% over the second best-performing score for the mel-scaled spectrogram and 2.33% for MFCC spectrogram in terms of validation accuracy, whereas, for the average accuracy, we obtain an increase of 0.79% and 0.44%, respectively, compared to the following best results.

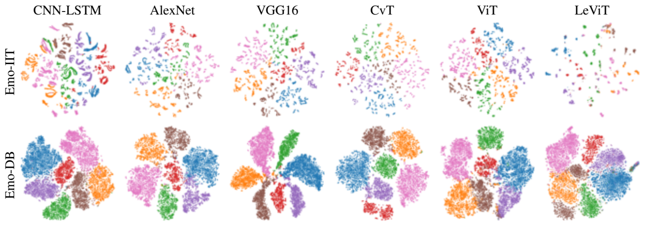


**Figure 3**: t-SNE plots for network architectures with a linear spectrogram on Emo-IIT and Emo-DB datasets. The blue color is anger, orange is boredom, green is disgust, red is fear, purple is happiness, brown is neutral, and pink is sadness. The higher density on Emo-DB is due to a more extensive test dataset, 1487s vs. 649s.

## 5.2 Adversarial Results

**Adversarial Attack.** Before improving the robustness of our networks, we assess their default resistance performance to adversarial attacks. Table 2a shows that for all models, except ViT and LeViT, the EOT+PGD algorithm is the most effective in adversarial attacks, with a minimum attack average success rate of 82.03% and a maximum of 99.62%. The LeViT network is most vulnerable to the MI-FGSM algorithm, with a 99.97% success rate and a 99.51% average success rate. Also, it is the least resistant to attacks overall, with a minimum adversarial success rate of 90.88%. The other models are most resistant to MI-FGSM attacks, with 76.64% attack average success for CNN-LSTM, 80.83% for AlexNet, 75.54% for VGG16, 79.37% for CvT, and 63.36% for ViT. The most invulnerable network without prior adversarial training is CNN-LSTM.

**Adversarial Defense.** Adversarial examples used as data augmentations (see No Atk. in Table 2a) determine an average accuracy improvement over undefended approaches of 0.65% for CNN-LSTM, 1.29% for VGG16, 0.34% for CvT, and 1.19% for ViT. On the other hand, the augmentations decrease the performance of AlexNet and LeVit by 0.38% and 2.22%. We obtained our best-performing model on the Emo-IIT dataset with CvT architecture, scoring 89.81% validation weighted accuracy and 98.09% average weighted accuracy, followed by VGG16 with 88.02% validation weighted accuracy and 97.78% average weighted accuracy. We find that CNN-LSTM architecture works better with BIM augmentations, VGG16 with EOT+PGD, CvT with PGD, and ViT with MI-FGSM.

For the defended adversarial attacks, Table 2a reveals consistent success rates for network architectures and adversarial algorithms. EOT+PGD is the most effective algorithm on AlexNet, VGG16, CvT, and LeViT, while FGSM takes the lead when attacking the CNN-LSTM and ViT architectures. Quantitatively, the VGG16 architecture is the most resistant to attacks. Comparing the defense capability, we obtain 31.73% average defense improvement for CNN-LSTM, 90.13% for AlexNet, 90.93% for VGG16, 13.26% for CvT, 30.88% for ViT, 4.66% for LeViT. We find convolution-based networks are better defended when adversarial training is performed, compared to transformer-based architectures, which achieve a maximum of 30.88% defense capability.

**Adversarial Perturbation Factor Variation.** Table 2b shows the best white-box attacks against AlexNet of 71.99% attack success for EOT+PGD with $\epsilon = 196/255$, followed by the BIM algorithm resulting in 11.10% attack success, and 10.96% for FGSM, both with $\epsilon = 8/255$. The weakest attacks are PGD with 10.86% for $\epsilon = 32/255$ and MI-FGSM with 10.84% for $\epsilon = 128/255$.

The $\epsilon$ hyperparameter variation of black-box attacks with the Pixle algorithm results in improvements of average weighted accuracy overall, with the highest attack success of 1.22% for FGSM-defended models and 0.18% for MI-FGSM models. The other algorithms used in adversarial training prove better resistance, showing top improvements of 0.80%, 0.60%, and 0.96% for BIM, PGD, and EOT+PGD. However, we saw improvements for FGSM and MI-FGSM of 0.38% and 0.23% as well. An $\epsilon > 0.5$ determines almost 100% attack success rate (all algorithms except BIM with 99.99%). We obtain the lowest numbers with $\epsilon = 8/255$ for FGSM models (2.72% attack success), MI-FGSM (2.39%), BIM (1.66%), PGD (1.90%), and EOT+PGD (1.68%). Moreover, an $\epsilon > 0.25$ determines attack results bigger than 95%, except EOT+PGD with 94.77%. The attack success correlation with the increase of $\epsilon$ is justified for the Square algorithm by the large-patch attacks.

## 5.3 Discussions

**Saliency Maps.** We use the visual representations of the deep neural network's focus in classifying the sentence "Merg la munte" (eng. "I am going to the mountain") from the Emo-IIT dataset on all emotions. Based on Figure 4, we identify consistent results for AlexNet trained with different augmentations. The focus regions on various test examples for these models are similar, and all the spectrograms are correctly classified. In contrast, the CvT models show no saliency map consistency for the same spectrograms. We find that there are classes featuring regions with more emphasis on the frequency levels, such as anger and boredom, with greater coverage of the low frequencies for the entire sentence. Moreover, other classes, such as

**Table 2**: Adversarial attack and defense results. Attack success (succ.) rates are computed based on average weighted accuracy. A negative attack success rate represents an accuracy gain on the Emo-IIT test data obtained by applying adversarial noise over the unattacked test data. The adversarial intensity $\epsilon$ varies for adversarial data augmentation and white-box, Pixle, and Square attacks. Higher values (↑) represent better performance. Lower numbers (↓) represent better model defense against adversarial attacks. Bold highlights the highest value.

(a) Adversarial attacks on regular and defended models.

| Model | Attack | Undefended (%) | | | | Defended (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Valid. Acc. ↑ | Avg. Acc. ↑ | Succ. ↓ | Avg. Succ. ↓ | Valid. Acc. ↑ | | Avg. Acc. ↑ | | Succ. ↓ | Avg. Succ. ↓ |
| | | | | | | No Atk. | Atk. | No Atk. | Atk. | | |
| CNN-LSTM | - | **77.24** | **95.75** | - | - | - | - | - | - | - | - |
| | FGSM | 25.07 | 19.04 | 67.55 | 80.12 | 77.69 | 9.23 | 95.64 | 47.53 | **88.11** | **50.30** |
| | MI-FGSM | 24.47 | 22.37 | 68.32 | 76.64 | 78.62 | **48.25** | 96.02 | **90.34** | 38.62 | 5.91 |
| | BIM | 21.64 | 19.14 | 71.98 | 80.02 | **80.60** | 44.01 | **96.40** | 89.54 | 45.40 | 7.11 |
| | PGD | 21.51 | 19.23 | 72.16 | 79.91 | 78.75 | 45.44 | 96.05 | 89.70 | 42.29 | 6.61 |
| | EOT+PGD | 19.45 | 17.20 | **74.82** | **82.03** | 78.70 | 45.91 | 96.04 | 89.64 | 41.66 | 6.67 |
| AlexNet | - | **84.35** | **97.05** | - | - | - | - | - | - | - | - |
| | FGSM | 0.39 | 3.50 | 99.53 | 96.39 | 81.50 | **79.33** | 96.54 | **96.07** | 2.66 | 0.49 |
| | MI-FGSM | 0.66 | 18.61 | 99.22 | 80.83 | 81.85 | 40.40 | 96.63 | 88.91 | 50.64 | 7.99 |
| | BIM | 0.03 | 1.91 | 99.96 | 98.03 | 81.47 | 33.08 | 96.54 | 87.61 | 59.39 | 9.25 |
| | PGD | 0.03 | 1.91 | 99.96 | 98.03 | 81.01 | 34.11 | 96.46 | 87.78 | 57.90 | 9.00 |
| | EOT+PGD | 0.00 | 0.37 | **100.00** | **99.62** | **82.19** | 32.47 | **96.67** | 87.49 | **60.49** | **9.49** |
| VGG16 | - | **81.09** | **96.49** | - | - | - | - | - | - | - | - |
| | FGSM | 2.00 | 11.87 | 97.54 | 87.70 | 79.23 | **71.33** | 96.11 | 90.77 | 9.97 | 5.55 |
| | MI-FGSM | 1.12 | 23.60 | 98.61 | 75.54 | 83.29 | 59.61 | 94.58 | 97.24 | 28.44 | -2.82 |
| | BIM | 0.10 | 4.90 | 99.87 | 94.92 | **88.02** | 50.37 | 96.89 | 90.40 | 42.78 | 6.69 |
| | PGD | 0.10 | 4.90 | 99.87 | 94.92 | 85.21 | 48.25 | 92.45 | **97.59** | 43.38 | -5.56 |
| | EOT+PGD | 0.00 | 1.34 | **100.00** | **98.61** | 87.20 | 47.96 | **97.78** | 90.27 | **45.00** | **7.68** |
| CvT | - | **88.12** | **97.75** | - | - | - | - | - | - | - | - |
| | FGSM | 0.50 | 10.22 | 99.43 | 89.54 | 80.56 | 5.58 | 96.34 | 26.55 | 93.08 | 72.44 |
| | MI-FGSM | 0.78 | 20.17 | 99.12 | 79.37 | 87.08 | **52.62** | 97.57 | **91.01** | 39.58 | 6.72 |
| | BIM | 0.14 | 5.17 | 99.84 | 94.71 | 89.51 | 24.72 | 98.04 | 66.15 | 72.38 | 32.53 |
| | PGD | 0.14 | 5.17 | 99.84 | 94.71 | **89.81** | 26.17 | **98.09** | 74.29 | 70.86 | 24.26 |
| | EOT+PGD | 0.00 | 1.41 | **100.00** | **98.55** | 88.37 | 2.69 | 97.85 | 14.39 | **96.96** | **85.29** |
| ViT | - | **77.41** | **94.88** | - | - | - | - | - | - | - | - |
| | FGSM | 1.74 | 2.33 | **97.75** | **97.54** | 73.64 | 15.24 | 92.02 | 30.68 | **79.31** | **66.66** |
| | MI-FGSM | 14.62 | 34.77 | 81.12 | 63.36 | 79.33 | **34.43** | **96.07** | **77.48** | 56.60 | 19.35 |
| | BIM | 3.02 | 4.91 | 96.10 | 94.83 | 77.22 | 28.03 | 93.73 | 56.56 | 63.70 | 39.65 |
| | PGD | 3.02 | 4.91 | 96.10 | 94.83 | **80.81** | 25.27 | 95.85 | 57.39 | 68.72 | 40.13 |
| | EOT+PGD | 2.40 | 3.40 | 96.89 | 96.42 | 77.66 | 28.90 | 95.26 | 63.13 | 62.79 | 33.73 |
| LeViT | - | **81.88** | **96.60** | - | - | - | - | - | - | - | - |
| | FGSM | 5.52 | 8.81 | 93.26 | 90.88 | 56.13 | **32.77** | 82.16 | **50.18** | 41.62 | 38.92 |
| | MI-FGSM | 0.03 | 0.47 | **99.97** | **99.51** | 64.25 | 4.44 | 80.72 | 7.25 | 93.10 | 91.01 |
| | BIM | 0.24 | 0.62 | 99.70 | 99.36 | 69.62 | 3.21 | 88.34 | 4.48 | 95.38 | 94.93 |
| | PGD | 0.24 | 0.62 | 99.70 | 99.36 | **73.72** | 2.54 | **92.66** | 11.88 | 96.55 | 87.18 |
| | EOT+PGD | 1.39 | 2.27 | 98.30 | 97.65 | 61.07 | 1.54 | 86.85 | 4.47 | **97.48** | **94.85** |

(b) Adversarial intensity $\epsilon$ variation for AlexNet.

| Algorithm | $\epsilon/255$ | Aug. Acc. (%) ↑ | Attk. Success (%) ↓ | | |
|---|---|---|---|---|---|
| | | | White-box | Pixle | Square |
| FGSM | 8 | 96.96 | **10.96** | 0.96 | 2.72 |
| | 16 | 95.82 | 9.05 | -0.38 | 5.43 |
| | 32 | 97.20 | 5.39 | 1.05 | 41.74 |
| | 64 | 97.03 | 4.45 | 1.16 | 98.59 |
| | 128 | **97.31** | 3.82 | **1.22** | **100.00** |
| | 196 | 97.06 | 3.85 | 1.11 | **100.00** |
| MI-FGSM | 8 | 97.54 | 10.08 | 0.07 | 2.39 |
| | 16 | 97.25 | 10.38 | -0.21 | 5.10 |
| | 32 | 97.30 | 10.54 | -0.16 | 14.98 |
| | 64 | 97.27 | 10.65 | -0.23 | 97.74 |
| | 128 | **97.56** | **10.84** | **0.18** | **100.00** |
| | 196 | 97.53 | 10.54 | 0.08 | **100.00** |
| BIM | 8 | **97.22** | **11.10** | **-0.08** | 1.66 |
| | 16 | 97.01 | 11.04 | -0.28 | 3.94 |
| | 32 | 97.07 | 10.66 | -0.27 | 10.78 |
| | 64 | 96.91 | 10.78 | -0.49 | 95.24 |
| | 128 | 96.58 | 10.18 | -0.80 | 99.99 |
| | 196 | 96.92 | 10.71 | -0.39 | **100.00** |
| PGD | 8 | 96.99 | 10.53 | -0.19 | 1.90 |
| | 16 | 96.62 | 10.56 | -0.60 | 3.78 |
| | 32 | 96.71 | **10.86** | -0.44 | 10.41 |
| | 64 | 96.65 | 10.65 | -0.57 | 96.54 |
| | 128 | 96.83 | 10.80 | -0.37 | **100.00** |
| | 196 | **97.07** | 11.26 | **-0.11** | **100.00** |
| EOT+PGD | 8 | **97.09** | 11.00 | **-0.05** | 1.68 |
| | 16 | 96.69 | 10.23 | -0.40 | 3.30 |
| | 32 | 96.29 | 10.09 | -0.86 | 10.59 |
| | 64 | 96.64 | 10.62 | -0.48 | 94.77 |
| | 128 | 96.74 | 12.16 | -0.33 | **100.00** |
| | 196 | 96.22 | **71.99** | -0.96 | **100.00** |

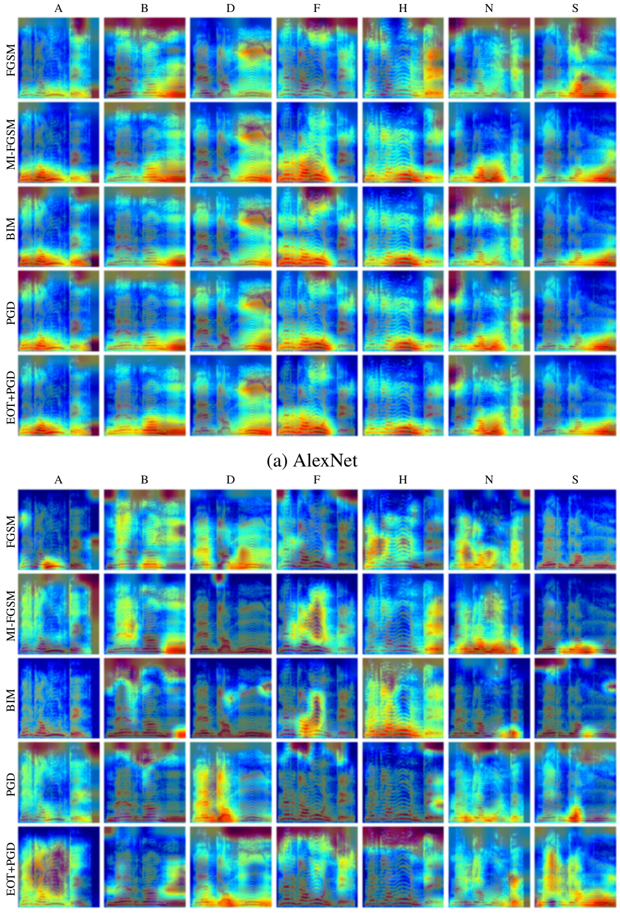
disgust, focus on specific time intervals and all frequency levels.

**Confusion Matrices.** We further investigate our data augmentation and defended adversarial attack results through the confusion matrices depicted in Figure 5. The highest confusion for AlexNet is in the classes {boredom, disgust, sadness}, with accuracy ranging 92.96-96.42%, while the classes {anger, fear, happiness, neutral} feature accuracy scores bigger than 97%. Similarly, the CvT's most significant confusion is for the {boredom, fear, sadness} classes, with the lowest accuracy of 90.30% for boredom on FGSM, followed by 94.06% accuracy for the EOT+PGD model. Compared to AlexNet, the classification is better with CvT, as shown in Tables 1, 2a.

In the offensive setup, the adversarial examples slightly affect AlexNet performance with a 5-15% drop in accuracy for almost every (class, defense algorithm) pair and with a few 2-3% performance drops in (anger, FGSM) and (happiness, FGSM) pairs. Moreover, some attacks have no success, resulting in accuracy improvements, for example, in (boredom, FGSM) case with 0.56% gain. Conversely, the CvT network presents significant attack vulnerabilities for FGSM and EOT+PGD, with high confusion overall with over 55% accuracy drops. BIM and PGD follow, causing 20-30% drops. Finally, the BIM-trained model well-defended the BIM attacks, resulting in only 2-10% accuracy reductions for individual classes.

**Insights.** Our results show that the linear and mel-scaled spectrograms perform best with the AlexNet and CvT networks. Although the convolutional transformers feature strong data augmentation performance, they are still poor in the SER task when using adversarial training as a defense mechanism. Furthermore, vision transformers are more prone to adversarial attack success on SER, and white-box attacks generally feature positive attack success rates. In contrast, black-box attacks can determine accuracy improvements at the pixel-level testing or complete adversarial attacks at the pixel patch level.

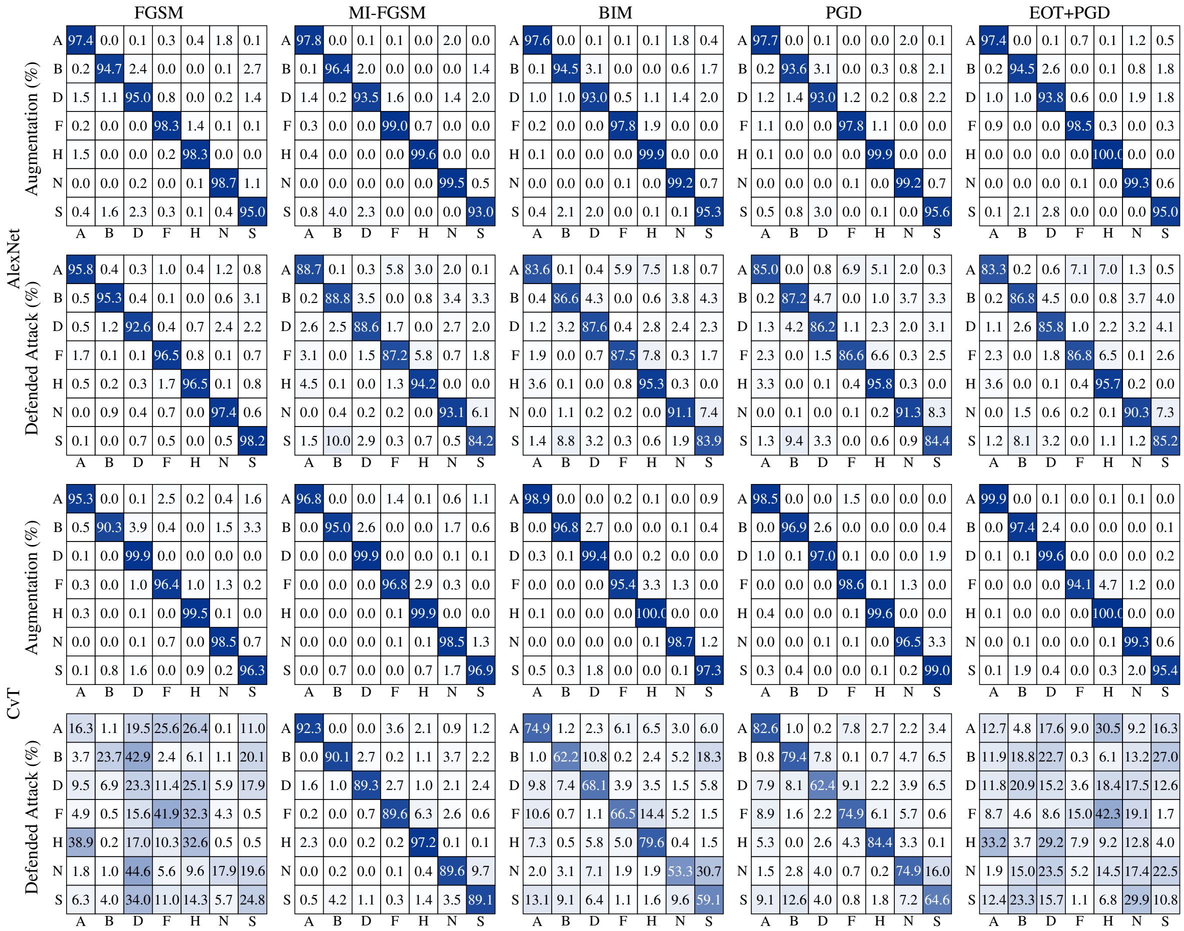
**Limitations.** Analyzing the experimental results, we notice that the vision transformers provide varying saliency maps that limit the conclusions on time, frequency, or decibels common focus regions. Moreover, the statistical stability will be further assessed as all our



(a) AlexNet

(b) CvT

**Figure 4**: Grad-CAM++ saliency maps for models trained with adversarial examples. The speech files used are "B303f{A, B, D, F, J, N, S}a.wav". The speaker and the sentence are the same for all files.



AlexNet — Augmentation (%)

| FGSM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 97.4 | 0.0 | 0.1 | 0.3 | 0.4 | 1.8 | 0.1 |
| B | 0.2 | 94.7 | 2.4 | 0.0 | 0.0 | 0.1 | 2.7 |
| D | 1.5 | 1.1 | 95.0 | 0.8 | 0.0 | 0.2 | 1.4 |
| F | 0.2 | 0.0 | 0.0 | 98.3 | 1.4 | 0.1 | 0.1 |
| H | 1.5 | 0.0 | 0.0 | 0.2 | 98.3 | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 0.2 | 0.0 | 0.1 | 98.7 | 1.1 |
| S | 0.4 | 1.6 | 2.3 | 0.3 | 0.1 | 0.4 | 95.0 |

| MI-FGSM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 97.8 | 0.0 | 0.1 | 0.1 | 0.0 | 2.0 | 0.0 |
| B | 0.1 | 96.4 | 2.0 | 0.0 | 0.0 | 0.0 | 1.4 |
| D | 1.4 | 0.2 | 93.5 | 1.6 | 0.0 | 1.4 | 2.0 |
| F | 0.3 | 0.0 | 0.0 | 99.0 | 0.7 | 0.0 | 0.0 |
| H | 0.4 | 0.0 | 0.0 | 0.0 | 99.6 | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 99.5 | 0.5 |
| S | 0.8 | 4.0 | 2.3 | 0.0 | 0.0 | 0.0 | 93.0 |

| BIM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 97.6 | 0.0 | 0.1 | 0.1 | 0.1 | 1.8 | 0.4 |
| B | 0.1 | 94.5 | 3.1 | 0.0 | 0.0 | 0.6 | 1.7 |
| D | 1.0 | 1.0 | 93.0 | 0.5 | 1.1 | 1.4 | 2.0 |
| F | 0.2 | 0.0 | 0.0 | 97.8 | 1.9 | 0.0 | 0.0 |
| H | 0.1 | 0.0 | 0.0 | 0.0 | 99.9 | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 99.2 | 0.7 |
| S | 0.4 | 2.1 | 2.0 | 0.0 | 0.1 | 0.1 | 95.3 |

| PGD | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 97.7 | 0.0 | 0.1 | 0.0 | 0.0 | 2.0 | 0.1 |
| B | 0.2 | 93.6 | 3.1 | 0.0 | 0.3 | 0.8 | 2.1 |
| D | 1.2 | 1.4 | 93.0 | 1.2 | 0.2 | 0.8 | 2.2 |
| F | 1.1 | 0.0 | 0.0 | 97.8 | 1.1 | 0.0 | 0.0 |
| H | 0.1 | 0.0 | 0.0 | 0.0 | 99.9 | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 99.2 | 0.7 |
| S | 0.5 | 0.8 | 3.0 | 0.0 | 0.1 | 0.0 | 95.6 |

| EOT+PGD | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 97.4 | 0.0 | 0.1 | 0.7 | 0.1 | 1.2 | 0.5 |
| B | 0.2 | 94.5 | 2.6 | 0.0 | 0.1 | 0.8 | 1.8 |
| D | 1.0 | 1.0 | 93.8 | 0.6 | 0.0 | 1.9 | 1.8 |
| F | 0.9 | 0.0 | 0.0 | 98.5 | 0.3 | 0.0 | 0.3 |
| H | 0.0 | 0.0 | 0.0 | 0.0 | 100.( | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 99.3 | 0.6 |
| S | 0.1 | 2.1 | 2.8 | 0.0 | 0.0 | 0.0 | 95.0 |

AlexNet — Defended Attack (%)

| FGSM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 95.8 | 0.4 | 0.3 | 1.0 | 0.4 | 1.2 | 0.8 |
| B | 0.5 | 95.3 | 0.4 | 0.1 | 0.0 | 0.6 | 3.1 |
| D | 0.5 | 1.2 | 92.6 | 0.4 | 0.7 | 2.4 | 2.2 |
| F | 1.7 | 0.1 | 0.1 | 96.5 | 0.8 | 0.1 | 0.7 |
| H | 0.5 | 0.2 | 0.3 | 1.7 | 96.5 | 0.1 | 0.8 |
| N | 0.0 | 0.9 | 0.4 | 0.7 | 0.0 | 97.4 | 0.6 |
| S | 0.1 | 0.0 | 0.7 | 0.5 | 0.0 | 0.5 | 98.2 |

| MI-FGSM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 88.7 | 0.1 | 0.3 | 5.8 | 3.0 | 2.0 | 0.1 |
| B | 0.2 | 88.8 | 3.5 | 0.0 | 0.8 | 3.4 | 3.3 |
| D | 2.6 | 2.5 | 88.6 | 1.7 | 0.0 | 2.7 | 2.0 |
| F | 3.1 | 0.0 | 1.5 | 87.2 | 5.8 | 0.7 | 1.8 |
| H | 4.5 | 0.1 | 0.0 | 1.3 | 94.2 | 0.0 | 0.0 |
| N | 0.0 | 0.4 | 0.2 | 0.2 | 0.0 | 93.1 | 6.1 |
| S | 1.5 | 10.0 | 2.9 | 0.3 | 0.7 | 0.5 | 84.2 |

| BIM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 83.6 | 0.1 | 0.4 | 5.9 | 7.5 | 1.8 | 0.7 |
| B | 0.4 | 86.6 | 4.3 | 0.0 | 0.6 | 3.8 | 4.3 |
| D | 1.2 | 3.2 | 87.6 | 0.4 | 2.8 | 2.4 | 2.3 |
| F | 1.9 | 0.0 | 0.7 | 87.5 | 7.8 | 0.3 | 1.7 |
| H | 3.6 | 0.1 | 0.0 | 0.8 | 95.3 | 0.3 | 0.0 |
| N | 0.0 | 1.1 | 0.2 | 0.2 | 0.0 | 91.1 | 7.4 |
| S | 1.4 | 8.8 | 3.2 | 0.3 | 0.6 | 1.9 | 83.9 |

| PGD | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 85.0 | 0.0 | 0.8 | 6.9 | 5.1 | 2.0 | 0.3 |
| B | 0.2 | 87.2 | 4.7 | 0.0 | 1.0 | 3.7 | 3.3 |
| D | 1.3 | 4.2 | 86.2 | 1.1 | 2.3 | 2.0 | 3.1 |
| F | 2.3 | 0.0 | 1.5 | 86.6 | 6.6 | 0.3 | 2.5 |
| H | 3.3 | 0.0 | 0.1 | 0.4 | 95.8 | 0.3 | 0.0 |
| N | 0.0 | 0.1 | 0.0 | 0.1 | 0.2 | 91.3 | 8.3 |
| S | 1.3 | 9.4 | 3.3 | 0.0 | 0.6 | 0.9 | 84.4 |

| EOT+PGD | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 83.3 | 0.2 | 0.6 | 7.1 | 7.0 | 1.3 | 0.5 |
| B | 0.2 | 86.8 | 4.5 | 0.0 | 0.8 | 3.7 | 4.0 |
| D | 1.1 | 2.6 | 85.8 | 1.0 | 2.2 | 3.2 | 4.1 |
| F | 2.3 | 0.0 | 1.8 | 86.8 | 6.5 | 0.1 | 2.6 |
| H | 3.6 | 0.0 | 0.1 | 0.4 | 95.7 | 0.2 | 0.0 |
| N | 0.0 | 1.5 | 0.6 | 0.2 | 0.1 | 90.3 | 7.3 |
| S | 1.2 | 8.1 | 3.2 | 0.0 | 1.1 | 1.2 | 85.2 |

CvT — Augmentation (%)

| FGSM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 95.3 | 0.0 | 0.1 | 2.5 | 0.2 | 0.4 | 1.6 |
| B | 0.5 | 90.3 | 3.9 | 0.4 | 0.0 | 1.5 | 3.3 |
| D | 0.1 | 0.0 | 99.9 | 0.0 | 0.0 | 0.0 | 0.0 |
| F | 0.3 | 0.0 | 1.0 | 96.4 | 1.0 | 1.3 | 0.2 |
| H | 0.3 | 0.0 | 0.1 | 0.0 | 99.5 | 0.1 | 0.0 |
| N | 0.0 | 0.0 | 0.1 | 0.7 | 0.0 | 98.5 | 0.7 |
| S | 0.1 | 0.8 | 1.6 | 0.0 | 0.9 | 0.2 | 96.3 |

| MI-FGSM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 96.8 | 0.0 | 0.0 | 1.4 | 0.1 | 0.6 | 1.1 |
| B | 0.0 | 95.0 | 2.6 | 0.0 | 0.0 | 1.7 | 0.6 |
| D | 0.0 | 0.0 | 99.9 | 0.0 | 0.0 | 0.1 | 0.1 |
| F | 0.0 | 0.0 | 0.0 | 96.8 | 2.9 | 0.3 | 0.0 |
| H | 0.0 | 0.0 | 0.0 | 0.1 | 99.9 | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 98.5 | 1.3 |
| S | 0.0 | 0.7 | 0.0 | 0.0 | 0.7 | 1.7 | 96.9 |

| BIM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 98.9 | 0.0 | 0.0 | 0.2 | 0.1 | 0.0 | 0.9 |
| B | 0.0 | 96.8 | 2.7 | 0.0 | 0.0 | 0.1 | 0.4 |
| D | 0.3 | 0.1 | 99.4 | 0.0 | 0.2 | 0.0 | 0.0 |
| F | 0.0 | 0.0 | 0.0 | 95.4 | 3.3 | 1.3 | 0.0 |
| H | 0.1 | 0.0 | 0.0 | 0.0 | 100.( | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 98.7 | 1.2 |
| S | 0.5 | 0.3 | 1.8 | 0.0 | 0.0 | 0.1 | 97.3 |

| PGD | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 98.5 | 0.0 | 0.0 | 1.5 | 0.0 | 0.0 | 0.0 |
| B | 0.0 | 96.9 | 2.6 | 0.0 | 0.0 | 0.0 | 0.4 |
| D | 1.0 | 0.1 | 97.0 | 0.1 | 0.0 | 0.0 | 1.9 |
| F | 0.0 | 0.0 | 0.0 | 98.6 | 0.1 | 1.3 | 0.0 |
| H | 0.4 | 0.0 | 0.0 | 0.1 | 99.6 | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 96.5 | 3.3 |
| S | 0.3 | 0.4 | 0.0 | 0.0 | 0.1 | 0.2 | 99.0 |

| EOT+PGD | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 99.9 | 0.0 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 |
| B | 0.0 | 97.4 | 2.4 | 0.0 | 0.0 | 0.0 | 0.1 |
| D | 0.1 | 0.1 | 99.6 | 0.0 | 0.0 | 0.0 | 0.2 |
| F | 0.0 | 0.0 | 0.0 | 94.1 | 4.7 | 1.2 | 0.0 |
| H | 0.1 | 0.0 | 0.0 | 0.0 | 100.( | 0.0 | 0.0 |
| N | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 99.3 | 0.6 |
| S | 0.1 | 1.9 | 0.4 | 0.0 | 0.3 | 2.0 | 95.4 |

CvT — Defended Attack (%)

| FGSM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 16.3 | 1.1 | 19.5 | 25.6 | 26.4 | 0.1 | 11.0 |
| B | 3.7 | 23.7 | 42.9 | 2.4 | 6.1 | 1.1 | 20.1 |
| D | 9.5 | 6.9 | 23.3 | 11.4 | 25.1 | 5.9 | 17.9 |
| F | 4.9 | 0.5 | 15.6 | 41.9 | 32.3 | 4.3 | 0.5 |
| H | 38.9 | 0.2 | 17.0 | 10.3 | 32.6 | 0.5 | 0.5 |
| N | 1.8 | 1.0 | 44.6 | 5.6 | 9.6 | 17.9 | 19.6 |
| S | 6.3 | 4.0 | 34.0 | 11.0 | 14.3 | 5.7 | 24.8 |

| MI-FGSM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 92.3 | 0.0 | 0.0 | 3.6 | 2.1 | 0.9 | 1.2 |
| B | 0.0 | 90.1 | 2.7 | 0.2 | 1.1 | 3.7 | 2.2 |
| D | 1.6 | 1.0 | 89.3 | 2.7 | 1.0 | 2.1 | 2.4 |
| F | 0.2 | 0.0 | 0.7 | 89.6 | 6.3 | 2.6 | 0.6 |
| H | 2.3 | 0.0 | 0.2 | 0.2 | 97.2 | 0.1 | 0.1 |
| N | 0.0 | 0.2 | 0.0 | 0.1 | 0.4 | 89.6 | 9.7 |
| S | 0.5 | 4.2 | 1.1 | 0.3 | 1.4 | 3.5 | 89.1 |

| BIM | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 74.9 | 1.2 | 2.3 | 6.1 | 6.5 | 3.0 | 6.0 |
| B | 1.0 | 62.2 | 10.8 | 0.2 | 2.4 | 5.2 | 18.3 |
| D | 9.8 | 7.4 | 68.1 | 3.9 | 3.5 | 1.5 | 5.8 |
| F | 10.6 | 0.7 | 1.1 | 66.5 | 14.4 | 5.2 | 1.5 |
| H | 7.3 | 0.5 | 5.8 | 5.0 | 79.6 | 0.4 | 1.5 |
| N | 2.0 | 3.1 | 7.1 | 1.9 | 1.9 | 53.3 | 30.7 |
| S | 13.1 | 9.1 | 6.4 | 1.1 | 1.6 | 9.6 | 59.1 |

| PGD | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 82.6 | 1.0 | 0.2 | 7.8 | 2.7 | 2.2 | 3.4 |
| B | 0.8 | 79.4 | 7.8 | 0.1 | 0.7 | 4.7 | 6.5 |
| D | 7.9 | 8.1 | 62.4 | 9.1 | 2.2 | 3.9 | 6.5 |
| F | 8.9 | 1.6 | 2.2 | 74.9 | 6.1 | 5.7 | 0.6 |
| H | 5.3 | 0.0 | 2.6 | 4.3 | 84.4 | 3.3 | 0.1 |
| N | 1.5 | 2.8 | 4.0 | 0.7 | 0.2 | 74.9 | 16.0 |
| S | 9.1 | 12.6 | 4.0 | 0.8 | 1.8 | 7.2 | 64.6 |

| EOT+PGD | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| A | 12.7 | 4.8 | 17.6 | 9.0 | 30.5 | 9.2 | 16.3 |
| B | 11.9 | 18.8 | 22.7 | 0.3 | 6.1 | 13.2 | 27.0 |
| D | 11.8 | 20.9 | 15.2 | 3.6 | 18.4 | 17.5 | 12.6 |
| F | 8.7 | 4.6 | 8.6 | 15.0 | 42.3 | 19.1 | 1.7 |
| H | 33.2 | 3.7 | 29.2 | 7.9 | 9.2 | 12.8 | 4.0 |
| N | 1.9 | 15.0 | 23.5 | 5.2 | 14.5 | 17.4 | 22.5 |
| S | 12.4 | 23.3 | 15.7 | 1.1 | 6.8 | 29.9 | 10.8 |

**Figure 5**: Confusion matrices for the AlexNet and CvT networks trained with adversarial examples for data augmentation and defended attacks. The horizontal axis represents predicted labels, and the vertical one shows true labels. The classes are anger (A), boredom (B), disgust (D), fear (F), happiness (H), neutral (N), and sadness (S). Values depicted in the matrices are expressed in percentage (%).

experiments were performed for a limited number of iterations, i.e., average results on three experimental iterations. Finally, the best training hyperparameters are subject to further fine-tuning for the SER task in low-resource languages such as Romanian.

## 6 Conclusions

This paper explores four types of spectrograms to feed six networks of different architectural complexity and approaches to the speech emotion recognition task. Our results show the relationship between the dataset and types of preprocessing. The adversarial algorithms employed as data augmentation techniques improve performance for all network architectures except AlexNet and LeViT. For attack and defense settings, we found that most architectures behave differently when adversarial algorithms attack them. The most robust algorithm that significantly breaks the model defense is EOT+PGD, with attack average success rates up to 94.85% for LeViT and 85.29% for CvT. Moreover, we notice a big difference in defense performance between convolution-based and transformer-based architectures. The latter provides poor accuracy even after employing adversarial training. Overall, the $\epsilon$ hyperparameter does not influence the performance linearly, but it is subject to parameter tuning to achieve better defense and attack performance.

Further work will comprise the adversarial attack robustness comparison between multiple low-resource languages, such as Persian and Greek. Finally, a comprehensive assessment of $\epsilon$ variation on various network architectures, as well as white and black-box adversarial algorithms is the subject of further investigation.

## Acknowledgements

This work was supported by the NUST POLITEHNICA Bucharest through the PubArt program and from a grant GNAC ARUT 2023.

## References

[1] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu. Convolutional neural networks for speech recognition. *ACM Trans. Audio Speech Lang. Process.*, 22(10):1533–1545, 2014.

[2] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501, 2020.

[3] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML Stockholm*, volume 80, pages 284–293. PMLR, 2018.

[4] A. Badshah, J. Ahmad, N. Rahim, and S. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service*. Institute of Electrical and Electronics Engineers Inc., 2017.

[5] W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th ICLR 2018, Vancouver, BC, Canada*, 2018.

[6] J. C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *INTERSPEECH*, pages 1517–1520. ISCA, 2005.

[8] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE WACV*, pages 839–847, 2018.

[9] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. In H. Ko and J. H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Korea*, pages 346–350. ISCA, 2022.

[10] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *4th ICLR 2016, San Juan, Puerto Rico*, 2016.

[11] J. Deng, R. Socher, L. Fei-Fei, W. Dong, K. Li, and L.-J. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE CVPR*, pages 248–255, 2009.

[12] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. pages 9185–9193, 2018. doi: 10.1109/CVPR.2018.00957.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, May 3-7*, 2021.

[14] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn.*, 44(3):572–587, 2011.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *3rd ICLR 2015, San Diego, CA, USA, Conference Track Proceedings*, 2015.

[16] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *2021 IEEE/CVF ICCV, Montreal, Canada*, pages 12239–12249, 2021.

[17] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2137–2146. PMLR, 2018.

[19] H. Kim. Torchattacks : A pytorch repository for adversarial attacks. arXiv preprint arXiv:2010.01950, 2020.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego,USA, May 7-9*, 2015.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012.

[22] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR, Toulon, 2017, Workshop Track Proceedings*, 2017.

[23] T. A. Lampert and S. E. O'Keefe. A survey of spectrogram track detection algorithms. *Applied Acoustics*, 71(2):87–100, 2010.

[24] 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*, 2, 2020.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

[26] W.-H. Liao, W.-Y. Chen, and Y.-C. Wu. On the robustness of cross-lingual speaker recognition using transformer-based approaches. In *2022 26th ICPR*, pages 366–371, 2022.

[27] X. Liu, Y. Li, C. Wu, and C.-J. Hsieh. Adv-BNN: Improved adversarial defense through robust bayesian neural network. In *International Conference on Learning Representations*, 2019.

[28] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *6th ICLR 2018, Vancouver, BC, Canada*, 2018.

[29] S. Mahmud, X. Lin, and J.-H. Kim. Interface for human machine interaction for assistant devices: A review. In *2020 10th Annual CCWC*, pages 0768–0773, 2020.

[30] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[31] H. Meng, T. Yan, F. Yuan, and H. Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7: 125868–125881, 2019.

[32] F. S. Monica, F. Monica, and Z. M. Dan. A new emotional corpus for the romanian language. In *2016 International Conference on Development and Application Systems (DAS)*, pages 260–263, 2016.

[33] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger. Does audio deepfake detection generalize?, 2022.

[34] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pages 506–519. ACM, 2017.

[35] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, 2019.

[36] J. Pomponi, S. Scardapane, and A. Uncini. Pixle: a fast and effective black-box attack based on rearranging pixels. *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2022.

[37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28492–28518. PMLR, 2023.

[38] S. Ramakrishnan and I. M. M. El Emary. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, 52(3):1467–1478, 2013.

[39] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, 1986.

[40] S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. In G. Kubin and Z. Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Austria*, pages 3465–3469, 2019.

[41] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4779–4783. IEEE, 2018.

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[43] J. Singh, L. B. Saheer, and O. Faust. Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, 20(6), 2023.

[44] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16*, 2014.

[45] D. Ungureanu, S.-A. Toma, I.-D. Filip, B.-C. Mocanu, I. Aciobăniței, B. Marghescu, T. Balan, M. Dascalu, I. Bica, and F. Pop. Odin112–ai-assisted emergency services in romania. *Applied Sciences*, 13(1), 2023.

[46] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS 2017*, pages 5998–6008, 2017.

[48] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. In *2021 IEEE/CVF707 ICCV*, pages 22–31, 2021.

[49] H. Xu, Y. Ma, H. Liu, D. Deb, H. Liu, J. Tang, and A. K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *Int. J. Autom. Comput.*, 17(2):151–178, 2020.

[50] Y. Xu, X. Zhong, A. Jimeno Yepes, and J. H. Lau. Grey-box adversarial attack and defence for sentiment classification. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the NAACL: Human Language Technologies*, Online, 2021. Association for Computational Linguistics.

[51] Y. Zhang, Y. Song, J. Liang, K. Bai, and Q. Yang. Two sides of the same coin: White-box and black-box attacks for transfer learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 2989–2997, New York, NY, USA, 2020. Association for Computing Machinery.

[52] J. Zhao, X. Mao, and L. Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

[53] R. S. Zimmermann. Comment on "Adv-BNN: Improved adversarial defense through robust bayesian neural network". arXiv preprint arXiv:1907.00895, 2019.