

A Systematic Evaluation of Adversarial Attacks against Speech Emotion Recognition Models

Nicolas Facchinetti¹, Federico Simonetta², and Stavros Ntalampiras^{1*}

¹Department of Computer Science, University of Milan, Italy

² Gran Sasso Science Institute, L'Aquila, Italy

*Corresponding author: stavros.ntalampiras@unimi.it

Abstract

Speech emotion recognition (SER) has been constantly gaining attention in recent years due to its potential applications in diverse fields and thanks to the possibilities offered by deep learning technologies. However, recent studies have shown that deep learning models can be vulnerable to adversarial attacks. In this paper, we systematically assess this problem by examining the impact of various adversarial white-box and black-box attacks on different languages and genders within the context of SER. We first propose a suitable methodology for audio data processing, feature extraction, and convolutional neural network long short-term memory (CNN-LSTM) architecture. The observed outcomes highlighted the considerable vulnerability of CNN-LSTM models to adversarial examples (AEs). In fact, all the considered adversarial attacks are able to considerably reduce the performance of the constructed models. Furthermore, when assessing the efficacy of the attacks, minor differences were noted between the languages analyzed as well as between male and female speech. In summary, this work contributes to the understanding of the robustness of CNN-LSTM models, particularly in SER scenarios, and the impact of AEs. Interestingly, our findings serve as a baseline for a) developing more robust algorithms for SER, b) designing more effective attacks, c) investigating possible defenses, d) improved understanding of the vocal differences between different languages and genders, and e) overall, enhancing our comprehension of the SER task.

1 Introduction

The exploration of automatic emotional state detection from vocal expressions has drawn considerable attention in the contemporary era, primarily due to its potential applicability in a broad spectrum of fields such as human-computer interaction, psychology, entertainment, and education [1–4]. The introduction of deep learning techniques has markedly improved the performance of speech emotion recognition (SER) models, fostering the development of numerous applications for public use [5]. However, recent studies have underscored the vulnerability of deep learning models to adversarial examples (AEs) – carefully crafted input samples designed to mislead the model into

producing erroneous predictions [6]. This susceptibility could trigger serious consequences in an SER context, especially in applications that are integral to safety. Evaluating the robustness of SER models is crucial, given their prospective use in areas such as affective computing, human-robot interaction, and mental health monitoring. By assessing and enhancing the resilience of these models, researchers can aid in the creation of reliable and trustworthy tools for these vital real-world applications.

Adversarial attacks represent a substantial menace to SER systems, resulting in the erroneous interpretation of a speaker’s emotional condition, which might lead to severe implications. The following are five instances that illustrate the aforementioned risk:

- In the context of customer service, an adversarial attack on an SER model might misread a customer’s irritation as joy, culminating in an unsuitable response and possibly a discontented customer.
- In the setting of mental health diagnosis, an adversarial attack on an SER model utilized to identify depression could lead it to falsely categorize a patient as being in good health, resulting in a faulty diagnosis and insufficient treatment.
- In the realm of entertainment, an adversarial attack on an SER model employed to modulate a virtual assistant’s tone might result in responses that are incongruous with the user’s expectations, leading to bewilderment and annoyance.
- In a security-oriented scenario, an adversarial attack on an SER model used for detecting deceit during a police inquiry might lead to the misclassification of a suspect’s truthful declarations as falsehoods, resulting in unwarranted allegations and potential erroneous detentions.
- In the situation of a job interview, an adversarial attack on an SER model employed to assess a candidate’s emotional aptitude might lead to the misinterpretation of a candidate’s anxiety as hostility, culminating in an incorrect evaluation and potentially overlooking a competent candidate.

In recent years a novel branch of scientific research has been studying the impact of AEs on SER tasks. Research in this area primarily aims to gauge the resilience of SER systems against various forms of attack, and to devise methodologies to enhance model performance in the face of such threats [7]. Despite these efforts, the subject matter remains largely unexplored, necessitating further investigation to gain a more comprehensive understanding of how different attack techniques could potentially impact system performance. The outcomes of such research could, for instance, highlight specific types of attacks that are exceptionally proficient at misleading the model, or indicate that the model exhibits greater robustness against attacks that alter certain input data features. Furthermore, the study may reveal that a specific language or gender is more susceptible to these attacks.

Although the impact of AEs on image models has been extensively studied, their application to SER models is lacking research. Furthermore, it is not possible to draw the same conclusions since the input data of image classification and SER models are only superficially similar. In fact, while

the success of convolutional neural networks (CNNs) and transformer-based architectures has been extended to the audio processing domain, the input samples used in the audio domain differ from those in the image domain in two main aspects: a) they often consist of sparse matrices, with most entries close to zero, and b) they are not easily segmentable, meaning that the same sound source (i.e., object) is spread across the matrix and is not contiguous as in the case of image segmentation. We believe that these differences warrant further investigation into the use of AEs in SER tasks.

This paper endeavors to address this gap in the existing body of knowledge by scrutinizing the effects of multiple adversarial attacks on various languages and genders in the SER context. It is of paramount importance to assess the robustness of emotion recognition models, comprehend their limitations, and create more resilient algorithms. This process could entail evaluating the model’s precision in the face of AEs that have been manipulated to trick the model into generating erroneous predictions. Within this framework, the primary contributions of this research are to:

- conduct an exhaustive analysis of the susceptibility of convolutional neural network-long short-term memory (CNN-LSTM) models to AE in SER;
- compare the performance of diverse attack categories;
- investigate potential disparities in the attack across three distinct languages and between male and female vocal samples.

Following an initial exploration of the scientific literature pertaining to principal techniques and methodologies in SER and adversarial machine learning, an optimal neural network model was identified to address the problem at hand. Given that there is no single model architecture that performs well across multiple languages in the literature [8, 9], we designed a model consisting of a fusion of CNN and LSTM, and is trained using log Mel-spectrograms derived from audio samples embodying diverse emotional states. The current paper focuses on multiple languages and scrutinizes the impacts of various adversarial attacks on speech data from both genders. To this end, three distinct datasets are utilized: EmoDB [10] for German, EMOVO [11] for Italian, and RAVDESS [12] for English. A conscious decision was made to design and educate our unique model to ensure maximum flexibility during experimentation, rather than depending on preexisting pretrained models that may not yield satisfactory outcomes across all languages.

Interestingly, a multitude of attacks were employed with the objective of assessing their influence on the established models. A broad spectrum of varied attack methodologies [13] was assessed, and, when feasible, diverse parameter sets were employed. The white-box attacks included were the fast gradient sign method (FGSM) [14], the basic iterative method (BIM) [15], DeepFool [16], the Jacobian-based saliency map attack (JSMA) [17], and Carlini and Wagner (C&W) [18]. For the black-box attacks, the One-Pixel Attack (PixelAttack) [19, 20] and the Boundary Attack (BoundaryAttack) [21] were utilized in our experimentation. Following comprehensive experimentation, we present detailed results that evaluate the efficacy of SER models when subjected to various attacks, taking into account language and gender factors.

1.1 Analysis of the Literature

1.1.1 Speech Emotion Recognition

SER represents the computational challenge of discerning a speaker’s emotional state by examining the acoustic properties of their speech signal [22]. Emotions, being a crucial component of human communication, are manifested through various speech facets including pitch, tempo, intensity, and spectral features. Despite the complexity and diversity in emotional expression through speech, recent breakthroughs in machine learning and deep learning methodologies have propelled substantial advancements in this domain, thereby stimulating active research interest in SER.

SER is a potent instrument, with its utility extending to a range of fields including virtual assistant development, emotion detection in customer service, and mental health surveillance. This paper provides a concise review of some pioneering studies and their respective application areas. Among the earliest researchers, Nakatsu et al. [23] explored the application of SER in an interactive movie system. This system not only allowed viewers to watch the narrative but also to engage with it, employing emotion recognition to facilitate spontaneous interactions among computer characters. In a different context, Petrushin et al. [24] concentrated on identifying emotional states in telephone call center dialogs. Here, understanding the caller’s mental state proved beneficial in decision support systems for tasks such as prioritizing voice messages, assigning suitable agents for responses, and categorizing voice mails based on the emotions expressed by the caller. Their study revealed anger as the most identifiable emotion. Further, France et al. [25] used acoustical characteristics as markers of depression and suicide risk, aiding therapists in comprehending their patients’ concealed emotions and overall mental state. Lastly, Schuller et al. [26] proposed a method that combined acoustic features and linguistic information in the automotive industry to enhance car ride safety by monitoring the driver’s mental state. Remarkably, this approach could trigger safety measures, potentially preventing accidents.

The research paper by [22] categorizes datasets for SER into three distinct types, based on the method of sample collection. These are simulated, seminatural, and natural speech datasets. Among these, the simulated datasets are the most prevalent. They are synthesized by trained speakers who articulate the same text, each time embodying a different emotion. Such datasets typically portray a standard set of emotions and, due to their acted nature, exhibit less noise and realism in comparison to datasets of natural speech [22]. The datasets utilized in the present study are all of the simulated kind.

One of the most frequently used datasets for SER tasks is the Berlin Database of Emotional Speech (EMO-DB) [10]. This dataset features ten actors, evenly split by gender, who simulate a range of emotions while uttering ten German sentences of varying lengths. Specifically, seven emotions (neutral, anger, fear, joy, sadness, disgust, and boredom) are represented across approximately 800 sentences, including 700 primary samples and some secondary versions. The recordings were conducted in an anechoic chamber, and the resultant material was subjected to an automated listening test. Each sentence was assessed by a panel of 20 listeners. Additionally, electroglottograms are provided to facilitate more precise extraction of prosodic and voice quality features.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [12] is a com-

prehensive, gender-balanced dataset. It comprises emotional speech and song recordings from 24 professional North American actors (12 female, 12 male), each contributing 104 sentences. This results in a total of 7,356 speech and 3,036 song samples, including both facial and vocal expressions. The speech component encapsulates a spectrum of emotions such as calm, happiness, sadness, anger, fear, surprise, and disgust, whereas the song component encompasses calm, happiness, sadness, anger, and fear. Each emotional expression is represented at two levels of intensity: normal and strong, complemented by a neutral expression. The validation process for RAVDESS involved two stages, with an initial group of 247 raters from North America followed by another group of 72 participants.

The EMOVO [11] dataset, on the other hand, is the inaugural emotional corpus tailored for the Italian language. It consists of recordings from six actors (three males and three females), each delivering 14 sentences that simulate seven emotions: disgust, fear, anger, joy, surprise, sadness, and neutral. These recordings, made using professional equipment in the Fondazione Ugo Bordoni laboratories, total 588, with each actor contributing approximately 10 minutes of material. This culminates in an overall database duration of one hour. The validation of the EMOVO dataset involved two distinct groups of 24 individuals, achieving an overall recognition accuracy of 80%.

In the realm of existing SER models, the CNN-LSTM models have, in recent years, consistently exhibited remarkable effectiveness, achieving unparalleled results as evidenced in multiple studies [8, 9, 27–29]. These models, characterized by their deep architecture, possess the ability to independently extract high-quality features from the data. Consequently, we have chosen to employ log-Mel spectrograms, a decision informed by their proven compatibility with this particular architecture in previous research [8, 9, 29–33].

1.1.2 Adversarial Machine Learning

The concept of “adversarial machine learning” encompasses a collection of methodologies devised for instigating malevolent attacks by manipulating models using accessible information. Typically, machine learning (ML) models are constructed based on a particular train/test set derived from an identical statistical distribution [34]. However, upon deployment, the model may be subjected to interference from an attacker who manipulates the system’s operation by introducing meticulously designed input data. Such inputs, termed as adversarial examples (AEs) [35], comprise legitimate inputs modified by the addition of minimal, often undetectable, perturbations. These perturbations are designed to deceive the system, thereby altering the anticipated outcomes by exploiting certain susceptibilities, all while being accurately classified by a human observer. The susceptibility of numerous ML models, including neural networks, to attacks instigated by minor modifications to the model’s input during testing, is an important concern. Biggio et al. [36] underscored this point by illustrating the dependency of an ML model’s success on its robustness against adversarial data. Their exemplar was a malware detection system for PDF files, which relied on a differentiable discriminant function.

The research explored two distinct scenarios. The first scenario involved an attacker possessing comprehensive knowledge of the target classifier, including the feature space, the model type, and

the trained model. Conversely, in the second scenario, the attacker’s knowledge was limited. The employed attack strategy hinged on the gradient descent walk of the classifier’s discriminant function $g(x)$, which was assumed to be differentiable, or an approximation thereof. The findings highlighted that both support vector machines (SVM) and neural networks could be successfully evaded, even when the adversary’s understanding of the system was minimal.

Regarding contemporary state-of-the-art deep neural networks, which demonstrate remarkable generalization in classification tasks, one would anticipate robustness against minor perturbations of the input signal. However, Szegedy et al. [35] discovered that even a negligible yet carefully tailored perturbation of an input image could alter the network’s prediction. Intriguingly, the error rate induced by these meticulously crafted examples surpassed that of examples perturbed with Gaussian noise, even though the average distortion was less.

1.1.3 Adversarial ML and SER Models

The pioneering scheme for generating AE in the context of linguistic applications was introduced by [37]. In their work, the authors focused on three paralinguistic tasks, including SER. Rather than applying perturbations to specific acoustic features, they opted to directly manipulate the raw waveform of an audio recording. The dataset utilized for their study was IEMOCAP, and the models of choice were WaveRNN and WaveCNN [37], both of which are considered to be state-of-the-art.

In the task of emotion recognition, the two models demonstrated similar performance levels: WaveRNN achieved an accuracy rate of 84%, while WaveCNN slightly surpassed it with an accuracy of 85%. The authors employed FGSM [14], detailed further in Attack Algorithms, as their chosen attack strategy. This method was applied twice, using various values of ϵ . When the perturbation factor was set to 0.015, there was a considerable increase in the emotion recognition error rate for both models: from 16% to 48% for WaveRNN, and from 15% to 42.5% for WaveCNN. Notably, these rates approached an upper bound error rate of 50% when only two classes were considered. The authors also observed that the AEs generated could be profitably transferred from WaveCNN to WaveRNN.

An important observation made by the authors was that the perturbations introduced through their approach were not only smaller, but also more effective than those achievable through an attack at the Mel-frequency cepstral coefficient (MFCC) feature level.

The inaugural black-box adversarial attack on SER systems is put forward by the authors in [38]. This attack, currently known as the real-world noise (RWN) attack, subtly manipulates the speech signal by incorporating minute, indiscernible noise. Upon experimental evaluation, the classification error rates were found to be 56.87% and 66.87% for the FAU-AIBO and IEMOCAP datasets, respectively. The authors further explore the potential of adversarial examples in enhancing model security through adversarial training. By integrating adversarially crafted examples into the training set, they managed to decrease the error rate by approximately 10 percentage points. Furthermore, the use of a generative adversarial network to generate examples yielded superior results. Specifically, the generator was programmed to alter adversarial examples with the aim of deceiving a network that differentiates between adversarial and genuine examples. Interestingly, they discovered that the

inclusion of a random noise layer did not benefit SER, a finding that contrasts with its impact on images.

The authors in [32] present a methodology for enhancing the resilience of a CNN based SER system against adversarial assaults. They employ the FGSM to generate adversarial instances from log Mel-spectrograms extracted from the DEMoS dataset. The robustness of three distinct models, namely a four-layer CNN, a ResNet model, and a VGG model, is assessed against these adversarial instances. The experimental results reveal a considerable decline in the unweighted average recall (UAR) of the models, from 0.8 to 0.2, with an increase in the ϵ parameter of FGSM.

To address this vulnerability, the authors propose three robustness-enhancing strategies. The initial strategy involves a data augmentation approach, which incorporates FGSM-generated instances into the training set. Subsequently, two adversarial training methodologies are suggested: vanilla and similarity-based. The former method proves effective in enhancing performance on authentic data compared to the conventional training approach, whereas the latter demonstrates superior efficacy in defending against adversarial attacks.

The need for preprocessing steps is evident when dealing with audio-type inputs, as it is crucial to extract certain features [39]. Unlike CNNs, which can directly operate on image pixels, this context requires initial extraction of signal features such as MFCCs or a log Mel-spectrogram [39]. The complexity increases when considering white-box attacks like FGSM, which utilize the gradient of the targeted audio relative to the input to calculate the optimal perturbation. Although the backpropagation method is efficiently applicable in image recognition due to the differentiability of all layers, the scenario becomes intricate for SER systems. The complexity arises from the commonly extracted features, such as the introduction of nonlinearity during the computation of MFCCs and nonlinearity in the output due to the usage of numerous LSTM units [38, 40, 41]. In light of empirical studies within this context, iterative methods have demonstrated superior effectiveness compared to single-step approaches [41].

In this study, we conducted a comprehensive examination of the impact of AEs on SER systems, encompassing a broad spectrum of elements, from the nature of the attacks to the spoken language and the gender of the speaker. In the spirit of ensuring complete reproducibility of our methodology and findings, the implementation has been made publicly accessible at https://github.com/LIMUNIMI/thesis_adversarial_ml_audio.

2 Materials and Methods

This section delineates the methodologies employed within the proposed framework for audio pattern recognition, which is dedicated to processing audio data and training SER models. It also provides a brief overview of the attacks utilized in the study.

2.1 Experimental Design

In order to assess adversarial attacks for SER models, a specific pipeline was developed. The objective was to observe the impact of different attacks on different languages, while also controlling

for the influence of speaker characteristics such as sex.

Three datasets were used, each representing a different language: RAVDESS (English), EmoDB (German), and EMOVO (Italian). To ensure consistency in the evaluation of attack effectiveness, the same model was kept fixed throughout the experiments. Since there is currently no existing model capable of performing SER in multiple languages, a custom model was developed for this study. The data underwent a rigorous cleaning and preprocessing to obtain log-Mel spectrograms. Labels that exhibited high correlation or were heavily under or over-represented in the datasets were removed. To increase the amount of training data, data augmentation techniques were applied.

A total of seven different CNN-LSTM architectures were designed and evaluated. The first model (\mathcal{M}_0) was used to determine the most effective normalization procedure, which was found to be simple standardization. This normalization procedure was then applied to all the remaining tests. The seven architectures were compared, and the best performing multilanguage model structure (\mathcal{M}_1) was identified. Further fine-tuning was conducted on the number of LSTM units and other hyperparameters of this model.

Finally, the obtained model and datasets were utilized to assess the impact of adversarial attacks from the Adversarial Robustness Toolbox (ART) Python library (online at <https://adversarial-robustness-toolbox.readthedocs.io/en/latest/guide/notebooks.html>).

Overall, this comprehensive methodology allowed for a thorough evaluation of adversarial attacks on SER models, taking into consideration different languages and controlling for speaker characteristics. The described workflow is shown in Figure 1.

2.2 Datasets

The EmoDB and EMOVO databases share a similar quantity of samples, in contrast to the substantially larger RAVDESS database, which is approximately 2.5 times greater in size. A notable variance is observed in the number of actors across these databases, as previously discussed in Analysis of the Literature.

In terms of audio duration distributions, the mean and 25, 50, and 75 percentiles demonstrate comparable statistics across the databases. However, large discrepancies are evident in the maximum duration and interfile duration. Given our objective to assess the impact of various attacks on SER systems through comparative analysis across different languages and genders, it is essential to maintain uniform training parameterization across all architectures, inclusive of sample duration. When examining the labels, it is observed that EmoDB and EMOVO incorporate 7 emotions, whereas RAVDESS includes 8. Further details concerning the extracted metadata can be found in Section S1 of the Supplementary Materials.

The datasets under consideration presented several challenges, which are addressed in the subsequent experimental setup. These challenges include: a) heterogeneity in the duration and sample rate of audio files, b) inconsistency in the number of classes across datasets, c) scarcity of data.

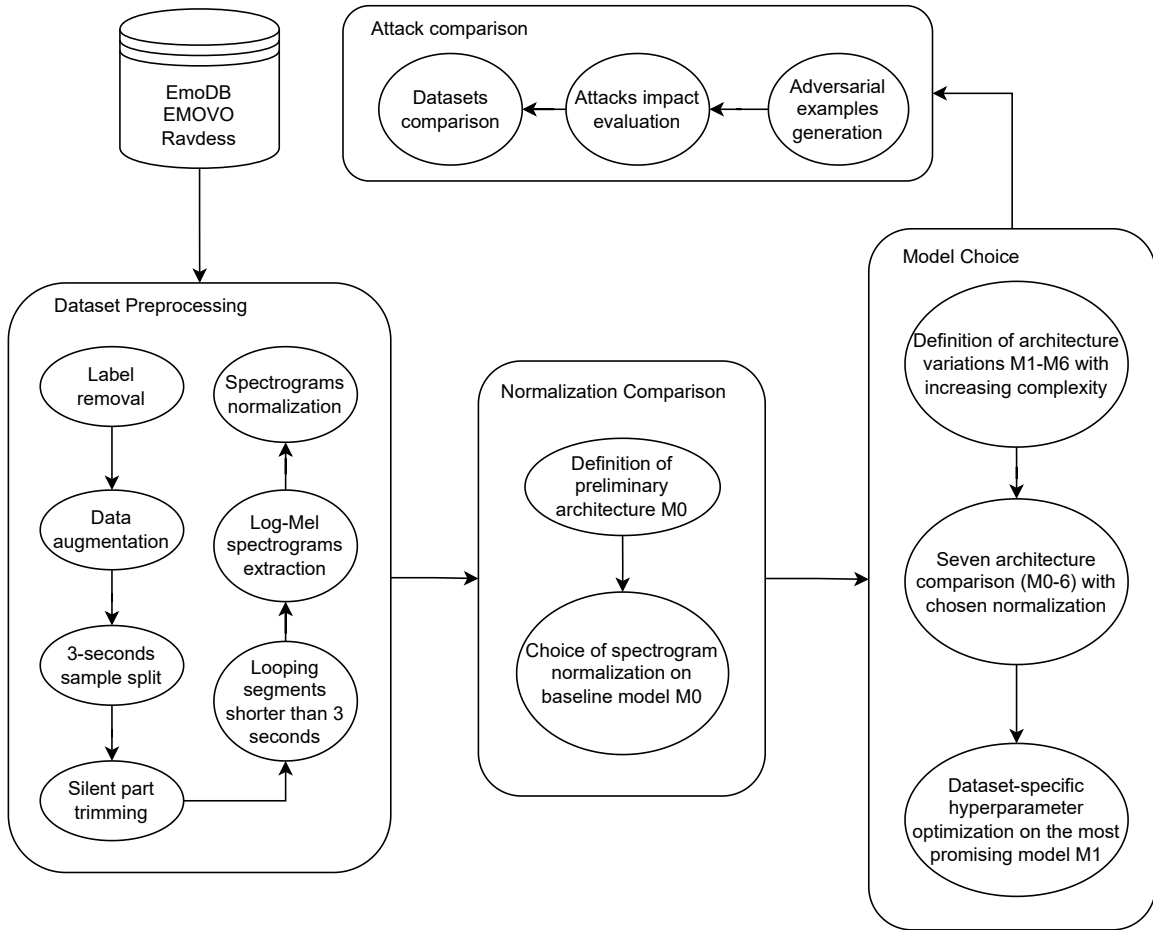


Figure 1: Flowchart of the proposed methodology for conducting the experiments.

2.2.1 Pre-processing

In order to obtain homogeneity of the samples, we applied some basic processing to the data. First, the datasets' samples are uniformly resampled at a frequency of 16,000 Hz and the silence at the commencement and termination of each signal is eliminated. Successively, segments with a duration less than 3 seconds are looped, while those exceeding this length are segmented into continuous, nonoverlapping 3-second intervals. This process of silence trimming is reiterated on the resulting segments to eliminate superfluous portions. Segments falling short of the 3-second standard are looped until they attain the requisite length. We finally computed the log-Mel spectrograms of the obtained audio excerpts, a choice motivated by prior research [9]. Log-Mel spectrograms were computed out using 128 Mel bands, a fast Fourier transformation window length of 368, and a hop size of 184. These parameters, corresponding to 23 ms and 11.5 ms, respectively, were set in line with a sampling rate of 16,000 Hz, as suggested by [42]. The spectrogram obtained was subsequently transformed to a logarithmic scale, leading to a 128×261 matrix saved for subsequent analysis. For illustrative purposes, an exemplar is provided in Figure 2.

Regarding the classes available in the datasets, since we are not interested in finding the optimal SER model, we chose to retain only five labels per dataset. We first computed log-Mel spectrograms of the audio excerpts obtained after segmentation, then, we used PCA and T-SNE to spot classes that were highly correlated. Moreover, other labels were discarded to improve the balance of the datasets. Specifically, the *fearful* and *angry* labels were omitted from EmoDB. As detailed in Section S1 of the Supplementary Materials, the *angry* label was the most prevalent, and its removal facilitated a more balanced dataset. For EMOVO, the excluded labels were *sad* and *angry*, while for RAVDESS, the *calm*, *neutral*, and *angry* labels were discarded. Similarly to the *angry* label in EmoDB, the *neutral* label in RAVDESS was disproportionately represented and its exclusion rectified this imbalance.

We approached the problem of the scarcity of data using data augmentation methods.

While generative adversarial networks have shown encouraging results in SER tasks [43–45], our approach is grounded in the application of less computationally intensive techniques. Data augmentation has the potential to considerably enhance the precision of a classifier and facilitate better model generalization to unobserved data, as the model becomes more resilient to the deformations applied [46]. According to [46], viable augmentation options encompass time stretching and pitch shifting. We employed an acceleration factor of $[0.75, 0.9, 1.1, 1.25]$ for time stretching. For pitch shifting, each step was configured to correspond to a semitone, with the values $[-3, -1.5, +1.5, +3]$ being considered.

The datasets are subjected to these processing steps, yielding eight supplementary augmented samples. To maintain a uniform duration of 3 seconds across all samples, identical procedures of division and repetition are subsequently applied to these newly generated samples.

2.2.2 Normalization

Once we obtained the augmented datasets, we were ready for continuing with subsequent phase as described in Experimental Design. However, one remaining step for preparing the data for neural models was the normalization of the log-Mel spectrograms. In the literature, various methods are

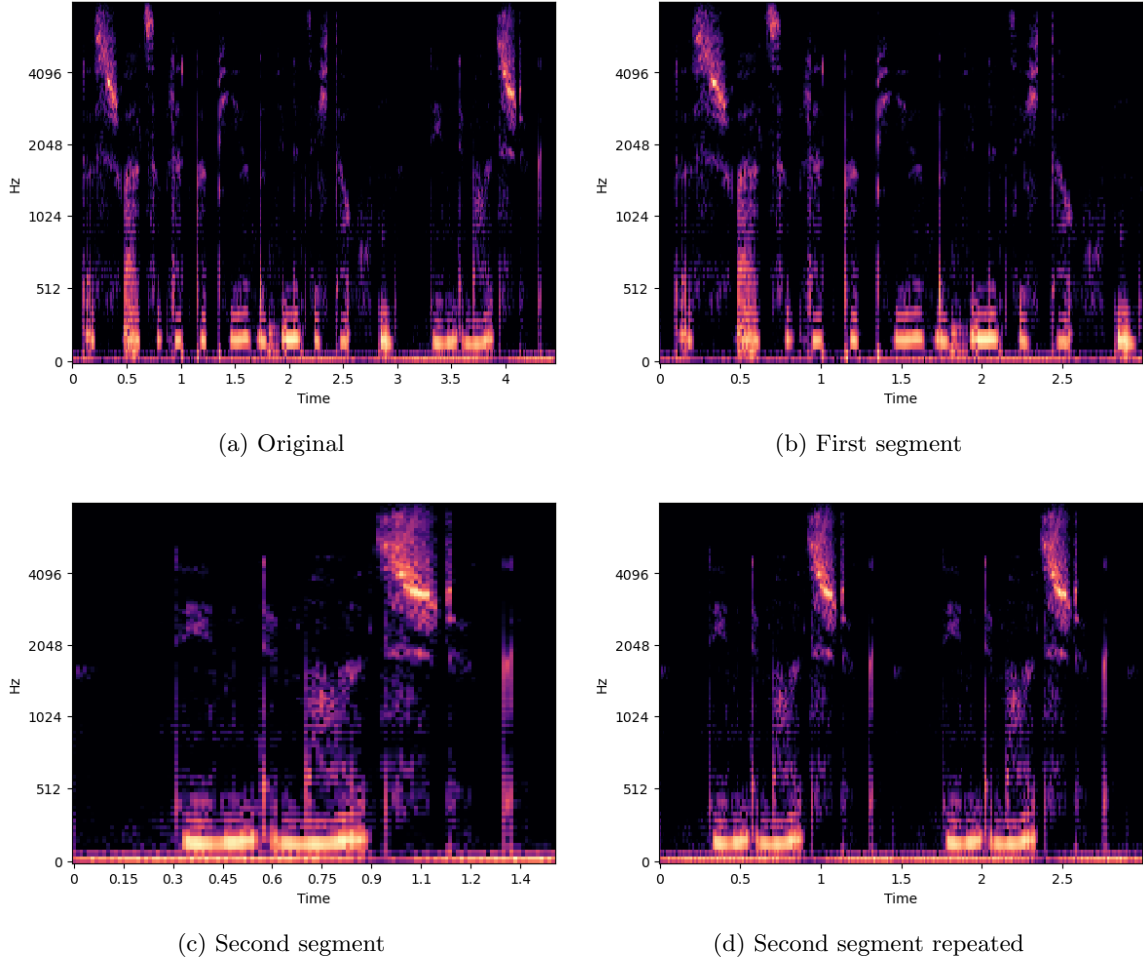


Figure 2: Example of split and repeat process on log Mel-spectrograms. Original log Mel-spectrogram (a), the sliced segments (b) and (c), and segment (c) repeated to 3 seconds (d).

adopted without a single methodology being more successful than the others. We were therefore interested in testing them. The objective of the normalization procedure is to transform the input values into a range suitable for neural learning, typically either $[0, 1]$ or $[-1, +1]$.

In the initial stages, four distinct transformations of the features are contemplated:

- **Original:** Utilizes the log Mel-spectrogram matrices devoid of any normalization, functioning as a baseline for assessing the practical benefits of the subsequent transformations.
- **NormSum:** Implements normalization by dividing each element of the matrix by the aggregate of all elements, thereby ensuring uniform sound energy across all samples. However, the values of each cell are small and approach zero.
- **NormMaxGlobal:** This method normalizes by dividing each matrix element by the maximum value across the dataset, ensuring all values fall within the $[0, 1]$ range while preserving the original proportions between the cells across varying spectrograms.

- **NormMaxLocal**: Normalization is achieved by scaling each element in a matrix through division by the maximum value within that specific spectrogram. This process ensures each matrix has a minimum value of 0 and a maximum value of 1, but the original proportions between cells are not preserved.

Beyond the aforementioned transformations, we opted to standardize each previously delineated version. This decision was informed by preliminary experiments that indicated a notable instability in the model’s learning process, evidenced by large fluctuations in the loss function across epochs. In this standardization phase, we initially transformed the matrices into arrays, subsequently standardized these, and then reshaped the data to its original form.

For assessing the best normalization strategy, we developed a rudimentary CNN-LSTM, designated as $\mathcal{M}0$, which demonstrated notable potential in preliminary investigations – see Models. The accuracy and its corresponding standard deviation, derived from each processing variant and dataset, are presented in Table 1 as per the definition provided in Results and Discussion.

Processing	EmoDB	EMOVO	RAVDESS	Average
Original	0.68 ± 0.04	0.35 ± 0.13	0.45 ± 0.1	0.50 ± 0.09
NormSum	0.46 ± 0.16	0.29 ± 0.10	0.21 ± 0	0.32 ± 0.09
NormMaxGlobal	0.79 ± 0	0.36 ± 0.20	0.21 ± 0	0.45 ± 0.07
NormMaxLocal	0.63 ± 0.07	0.48 ± 0.02	0.60 ± 0.03	0.57 ± 0.04
Original Standardized	0.71 ± 0.02	0.60 ± 0.07	0.62 ± 0.01	0.64 ± 0.04
NormSum Standardized	0.71 ± 0.01	0.35 ± 0.19	0.21 ± 0	0.42 ± 0.07
NormMaxGlobal Standardized	0.68 ± 0.02	0.24 ± 0.03	0.21 ± 0	0.38 ± 0.02
NormMaxLocal Standardized	0.68 ± 0.02	0.42 ± 0.15	0.58 ± 0.01	0.56 ± 0.06

Table 1: Mean accuracy \pm standard deviation over the three splits for each dataset and normalization type.

The assessment indicates that the use of *NormMaxGlobal* in the context of EmoDB and *Original Standardized* for EMOVO and RAVDESS, leads to enhanced accuracies as reflected in Table 1. Consistently, the standardized variant yields more reliable results, corroborating the preliminary experimental observations. Additionally, the loss function exhibits fewer temporal variations, fostering a steadier learning trajectory.

Although *NormMaxGlobal* exhibited superior performance with respect to accuracy and standard deviation on the EmoDB dataset, we opted to utilize *Original Standardized* consistently across all three instances, which maximizes the accuracy across the datasets on average. The standardization was applied to the datasets after the above-described preprocessing steps.

2.3 Models

2.3.1 Multi-dataset Architecture

We developed 7 model architectures to search for an optimal model across the three datasets that could be used for a fair assessment of the attacks. The base model $\mathcal{M}0$ incorporates three *Conv2D* layers, exhibiting an increase in filters (16, 32, 64) and a decrease in square kernel size, in a manner

akin to the models presented in [47, 48]. The activation function employed is *ReLU*, as suggested by [48]. Following each convolutional layer is a *MaxPooling* layer, with a pool size that varies ((from (4,4) initially to (2,2) subsequently)) and strides (initially 2, later 1), mirroring the approach in [47, 48]. To mitigate overfitting, expedite training, and facilitate the adoption of elevated learning rates, a *BatchNormalization* layer is positioned post the initial convolution [49]. The output from the CNN is flattened and transferred to an *LSTM* layer comprising three internal units, configured with an internal dropout of 0.2. The final output layer is a *Dense* layer with five units, employing a *Softmax* activation function to generate the probability for the five labels present in each dataset. A comprehensive delineation of the architecture is provided in Section S2 of the Supplementary Materials.

We then designed 6 other models, each exhibiting a progressive increase in parameter count and complexity. These models bear the nomenclature $\mathcal{M}1$, $\mathcal{M}2$, $\mathcal{M}3$, $\mathcal{M}4$, $\mathcal{M}5$, and $\mathcal{M}6$. The networks are sequentially arranged based on their parameter quantity, with slight variations in both CNN structure and LSTM internal unit count, while maintaining a foundational structure akin to $\mathcal{M}0$. In-depth information regarding the architectures is provided in Section S2 of the Supplementary Materials. Every convolutional layer uniformly employs the same kernel size, with the exception of models $\mathcal{M}3$ and $\mathcal{M}6$, which utilize more intricate architectures. The pooling operations exhibit a similar pattern, save for the reduced pool size in the initial pooling layer of models $\mathcal{M}3$ and $\mathcal{M}4$. In the case of $\mathcal{M}5$, there is a doubling of the filter count for each layer, whereas model $\mathcal{M}6$ opts for a quartet over a trio. As for the LSTM aspect, all models function in a unidirectional manner, barring $\mathcal{M}1$ which operates bidirectionally. Models $\mathcal{M}2$, $\mathcal{M}4$, and $\mathcal{M}6$ incorporate six units, in contrast to the remaining models, which utilize three.

The models undergo training with a batch size of 32 across 50 epochs, utilizing the *Adam* optimization algorithm with a learning rate of 0.001, in line with [50]’s approach to a similar CNN-LSTM architecture. The *categorical_crossentropy* loss function is employed. To mitigate overfitting, an *EarlyStopping* callback is introduced with a tolerance of 10 epochs, which observes the validation loss. Failing to observe an improvement over 10 epochs prompts the restoration of the weights corresponding to the optimal validation loss. Moreover, a *ReduceLROnPlateau* callback is implemented to decrease the learning rate upon observing no improvement in validation loss over six epochs. This strategy aims to prevent the model from straying from the ideal solution due to an excessive learning rate, while also promoting convergence by adopting smaller steps towards the cost function’s optimal solution with a diminished learning rate. The underlying theory is that as the model nears a sub-optimal solution with the current learning rate, it oscillates around the global minimum. Reducing the rate allows for smaller steps towards the cost function’s optimal solution. The validation loss serves as the metric for this callback, and the learning rate is reduced by a factor of 0.1 when there is no improvement in the validation loss for six epochs.

In the realm of SER tasks, a prevalent evaluation strategy is the leave-one-speaker-out (LOSO) cross-validation method, which designates each unique speaker as a test set. Nonetheless, this research deviates from the norm, opting for a conventional train/validation/test partition due to the inconsistency in the number of actors across datasets, which inevitably yields diverse test sets. Such disparity could potentially skew performance assessments when juxtaposing attacks on varying

languages and genders. In particular, the data was apportioned into three splits, with proportions of 64%, 16%, and 20%, respectively. To ensure a more reliable estimation of the performance metrics, this procedure was reiterated three times, each time employing a distinct random seed.

The architectures and datasets under consideration were evaluated using identical train/validation/test splits, consistent with the methodology employed in the preceding stage. The mean accuracies and corresponding standard deviations for all the splits, across all architectures and datasets, are tabulated in Table 2. The $\mathcal{M}1$ architecture emerges as the most potent, demonstrating satisfactory performance across all three datasets. Compared to $\mathcal{M}0$, the accuracy exhibits a substantial enhancement: a gain of +0.12 points for EmoDB, +0.11 for EMOVO, and +0.23 for RAVDESS. Concurrently, a notable reduction in the standard deviation signifies a robust generalization capability. The bidirectional configuration’s efficacy is also evident, as it surpasses $\mathcal{M}2$ —an architecture with identical CNN and output shape post-LSTM, but with six unidirectional units.

Architecture	EmoDB	EMOVO	RAVDESS	Mean
$\mathcal{M}0$	0.71 ± 0.05	0.59 ± 0.03	0.53 ± 0.11	0.61 ± 0.06
$\mathcal{M}1$	0.83 ± 0.00	0.70 ± 0.03	0.76 ± 0.01	0.76 ± 0.01
$\mathcal{M}2$	0.78 ± 0.01	0.68 ± 0.03	0.70 ± 0.02	0.72 ± 0.02
$\mathcal{M}3$	0.68 ± 0.05	0.39 ± 0.14	0.39 ± 0.13	0.49 ± 0.10
$\mathcal{M}4$	0.76 ± 0.01	0.66 ± 0.03	0.45 ± 0.15	0.62 ± 0.06
$\mathcal{M}5$	0.58 ± 0.05	0.25 ± 0.03	0.26 ± 0.04	0.36 ± 0.04
$\mathcal{M}6$	0.32 ± 0.06	0.24 ± 0.03	0.23 ± 0.03	0.26 ± 0.04

Table 2: Mean accuracy and standard deviation over the three splits for each dataset and model architecture.

Furthermore, the convergence across all splits for each dataset is a shared characteristic of both $\mathcal{M}1$ and $\mathcal{M}2$, a trait not observed in other configurations. A noteworthy observation is the inverse relationship between model complexity and accuracy, implying an optimal parameter count in $\mathcal{M}1$ for the given training set size. Intriguingly, the application of more intricate or deeper networks does not confer any advantage for RAVDESS, despite its larger number of examples. Once more, EmoDB consistently outperforms, reinforcing the notion that it represents a less complex task.

Consequently, it is cogent to persist in the optimization of hyperparameters for $\mathcal{M}1$, with its architecture succinctly revisited in Section S2 of the Supplementary Materials.

2.3.2 Hyperparameter Optimization

The final stage in defining the model architecture involves hyperparameter tuning to augment the model’s performance on the respective datasets. The hyperparameter optimization was split into two parts: first, we optimized the number of LSTM units, then the remaining hyperparameters that did not impact the overall model architecture. For this, we used the Hyperband tuner [51] with categorical cross-entropy loss.

Prior studies indicated an enhancement in performance upon augmenting the LSTM layer’s output quantity, as evidenced by the 3 units in $\mathcal{M}0$ and the 6 units in both $\mathcal{M}1$ and $\mathcal{M}2$. Consequently, additional investigations were undertaken on the $\mathcal{M}1$ model to ascertain the ideal quantity of LSTM

units. In particular, we examined multiple bidirectional configurations, encompassing 4, 8, 16, 32, 64, 128, 256, 512, and 1,024 units.

Typically, the model’s accuracy is enhanced by augmenting the quantity of LSTM units, although this association does not consistently apply to the final four values. To illustrate, the configurations yielding the most superior outcomes, in descending order, are as follows:

- For EmoDB, the top four configurations encompass 512, 256, 128, and 1,024 units;
- For EMOVO, the top four configurations comprise 512, 256, 1024, and 128 units;
- For RAVDESS, the top four configurations consist of 1,024, 256, 512, and 128 units.

Consequently, the LSTM layer was equipped with 256 units for the following reasons:

1. It consistently delivered the second-highest performance across all three scenarios;
2. The performance disparity between the top configuration and this one is negligible;
3. It engenders a more streamlined model, thereby mitigating the potential for overfitting.

For additional insights pertaining to the tuning of model $\mathcal{M}1$, please refer to Section S3 in the Supplementary Materials.

Since the number of internal units in the LSTM layer has considerably increased, an additional dropout layer was added after the flatten layer to prevent possible overfitting. Figure 3 summarizes the final architecture of the model.

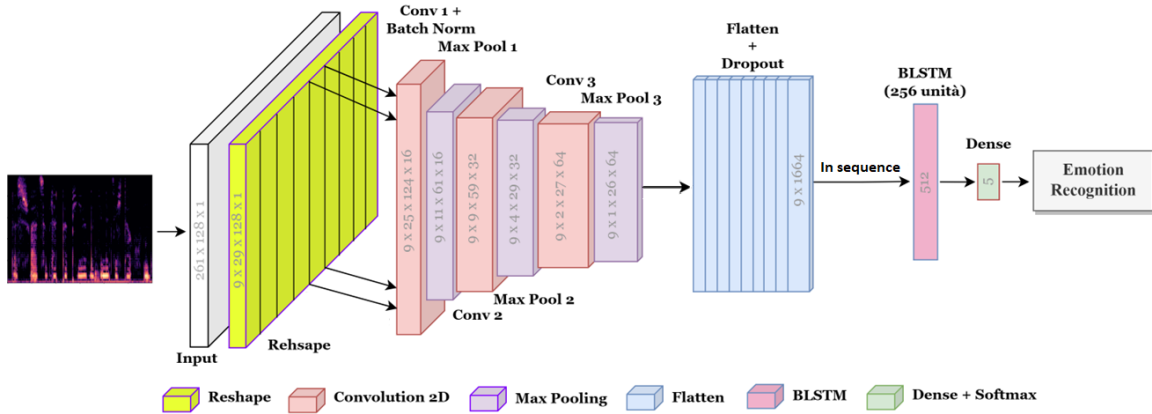


Figure 3: Architecture of the optimized CNN-LSTM model.

The second part of the hyperparameter optimization consisted of the fine-tuning of the dropout layer’s probability, the internal dropout probability of the BLSTM layer, the initial learning rate for the Adam optimizer, and the batch size.

The tuning of the latter two parameters, namely the learning rate and batch size, is of paramount importance to avert overfitting. These parameters considerably influence the model’s performance, convergence, and stability. They are intricately linked to the problem at hand, the input data, and the model’s architecture.

The hyperparameter tuning process has notably enhanced the performance, particularly for larger datasets like EMOVO and RAVDESS. For instance, the loss function showed considerable reductions:

- In the EmoDB dataset, the loss minimally decreased from 0.321270 to 0.312610, which is a reduction of 0.00866 or 2.69%;
- In the EMOVO dataset, the loss substantially decreased from 0.428224 to 0.344954, which is a reduction of 0.08327 or 19.44%;
- In the RAVDESS dataset, the loss remarkably decreased from 0.366705 to 0.281702, which is a reduction of 0.08501 or 23.18%.

This enhancement in performance through hyperparameter tuning has particularly allowed for a more effective utilization of datasets with a larger volume of data. Additional details regarding the optimization process can be referred to in Section S3 in the Supplementary Materials.

The final accuracies of the models are presented in Table 3.

Data	EmoDB	EMOVO	RAVDESS
All	0.909	0.872	0.911
Male	0.895	0.893	0.911
Female	0.918	0.852	0.911

Table 3: Accuracy on the original test set of the models that will be attacked.

2.4 Attack Algorithms

Our experimental study encompasses seven nontargeted algorithms: five white-box attacks and two black-box attacks. The white-box attacks under consideration are the fast gradient sign method (FGSM), the basic iterative method (BIM), DeepFool, the Jacobian-based saliency map attack (JSMA), and Carlini and Wagner (C&W). The black-box attacks are the One-Pixel Attack (PixelAttack) and the Boundary Attack (BoundaryAttack).

2.4.1 FGSM

The fast gradient sign method (FGSM) [14], a well-known and straightforward adversarial threat generation technique, primarily exploits the L_∞ distance metric. The method manipulates machine learning models by introducing minimal perturbations into the input data, thereby inducing the model to generate erroneous predictions. Specifically designed to exploit the learning mechanism of neural networks, FGSM utilizes the gradient of the loss with respect to the input data to augment the input data in a way that maximizes the loss. This is achieved by adding a noise vector, derived from the sign of the gradient, to the input data [48]. The following equation typically illustrates the generation of an adversarial example via this method:

$$X_{adv} = X + \epsilon * \text{sign}(\nabla_X J(X, y)). \quad (1)$$

Given an input X and its corresponding label y , the loss function $J(\cdot)$, and a noise magnitude parameter ϵ , the authors demonstrated the potential for successful adversarial attacks. The parameter ϵ is carefully chosen to balance two conflicting requirements: it must be sufficiently small to render the perturbations imperceptible to humans, yet large enough to mislead the model into making erroneous predictions. The efficacy of such attacks is contingent upon the model's gradient strength, the magnitude of the added noise, and the model's complexity.

The authors' experiments on the ImageNet dataset using the GoogLeNet CNN [14], revealed the ease with which highly effective adversarial examples could be generated. The AEs, created using the fast method, maintained a similar accuracy level until $\epsilon = 32$. Beyond this point, the accuracy gradually declined to nearly zero as ϵ increased to 128 [52]. This phenomenon can be attributed to the fact that FGSM adds noise scaled by ϵ to each image. Consequently, utilizing higher ϵ values effectively obliterates the image content, rendering it unrecognizable to humans.

2.5 BIM

The Basic Iterative Method (BIM) [15], an extension of FGSM utilizing the L_∞ distance metric, employs multiple iterations with a minimal step size. The BIM attack commences with an initial input image, following which the adversary calculates the gradient of the model's loss function relative to this image. Subsequently, the image is incrementally adjusted in the direction of this gradient. This iterative procedure continues either for a predetermined number of iterations or until the attainment of the desired output. Through continuous input modifications based on the model's gradient, the adversary can gradually manipulate the image to induce an erroneous prediction from the model. Specifically, during each iteration, the pixel values of the input image are confined to ensure their location within the ϵ -neighborhood of the original image.

The recursive function used to produce an AE from an input image X is the following:

$$X_0^{adv} = X, X_{N+1}^{adv} = Clip_{X,\epsilon}(X_N^{adv} + \alpha * sign(\nabla_x J(X_N^{adv}, y))). \quad (2)$$

The function $Clip_{X,\epsilon}(X')$ performs per-pixel clipping on the image X' , resulting in an L_∞ ϵ -neighborhood of the original image X . The label associated with X is represented by y , while the loss function and step size are denoted by $J()$ and ϵ , respectively. The authors empirically determined optimal values, setting $\alpha = 1$, which modifies each pixel value by 1 at every step. The number of iterations was chosen heuristically as $\min(\epsilon + 4, 1.25\epsilon)$, a balance ensuring that the AE would reach the boundary of the L_∞ ϵ -neighborhood while maintaining a manageable computational cost for experiments.

The experimental results [15] reveal that the iterative method induces subtler perturbations compared to FGSM, maintaining the integrity of the image even at high ϵ values. Concurrently, it confounds the classifier at a higher rate. Specifically, BIM generates superior AE for $\epsilon < 48$. Beyond this threshold, however, its performance plateaus and no further improvements are observed.

2.5.1 DeepFool

DeepFool (DF) [16] is an adversarial attack method that crafts AEs by iteratively computing the minimal L_2 perturbations. This iterative process seeks to ascertain the shortest distance from the original input to the decision boundary of the threat model. The underlying premise of the DeepFool technique is the local approximation of highly nonlinear deep neural networks by linear decision boundaries. This assumption allows the authors to analytically formulate the optimal solution to this simplified problem and subsequently construct the AE. Given that neural networks are not strictly linear, the algorithm incrementally moves towards the derived solution, repeating the process until a genuine AE is discovered. The DeepFool algorithm utilizes the resulting gradient to determine the optimal direction and magnitude of the perturbations necessary to induce a misclassification by the model. For any differentiable classifier, DeepFool presumes that f is linear around x'_t and iteratively computes the perturbation r_t :

$$\arg \min_{r_t} \|r_t\|_2 \text{ subject to } f(x'_t) + \nabla f(x'_t)^T r_t = 0. \quad (3)$$

The algorithm, in every iteration, computes the gradient of the decision function relative to the input data. Subsequently, it determines the least perturbation necessary to transition the input data point beyond the decision boundary into the subsequent class. This procedure is iteratively executed until the model misclassifies the input data point.

It has been demonstrated by the authors that the proposed technique effectively deceives advanced image recognition systems. Furthermore, the perturbation instigated by DeepFool is found to be less than that of FGSM across multiple benchmark datasets [16].

2.5.2 JSMA

The Jacobian Saliency Map Attack (JSMA) [17] is an effective adversarial methodology employing a greedy algorithm. This approach leverages L_0 distances to generate AEs by iteratively modifying individual pixels. The gradient of the loss, with respect to every input component, is exploited to identify key pixels and their corresponding perturbations. This is facilitated by a saliency map, which pinpoints the input features of importance to the adversary's objectives. The saliency map is derived from the forward derivative (Jacobian) of the function that a deep neural network (DNN) has learned. The procedure of the JSMA attack can be encapsulated as follows:

- Calculate the Jacobian matrix pertaining to the model's output with respect to the input image;
- Generate the saliency map;
- Select a target class, denoted as l , for the AE;
- From the saliency map, pinpoint the most influential features that augment the likelihood of class l while simultaneously diminishing the likelihood of the original class;

- Adjust the aforementioned features by a specified parameter θ , to create an AE that is erroneously classified as l ;
- Iterate over steps 2-5 until the AE is successfully produced.

The JSMA is a notably potent technique for the creation of AEs with minimal perturbations, which are often challenging to identify. The original work by Papernot et al. [17] provides a comprehensive explanation of the precise formulation employed, which we recommend for interested readers.

The authors demonstrate that this algorithm can consistently generate samples that, while perceived as correctly classified by human subjects, are misclassified by a DNN towards specific targets. This is achieved with an impressive adversarial success rate of 97% and an average modification of merely 4.02% of the input features per sample [17].

Despite its effectiveness, the JSMA technique may not always be the most efficient choice due to its computational cost. Furthermore, its performance may vary compared to other attack methods under certain circumstances.

2.5.3 C&W

The optimization-based attacks by Carlini and Wagner (C&W) [18] are capable of generating AE measured in L_0 , L_2 , and L_∞ norms. These attacks modify the objective and the primary constraint of the AE generation optimization problem, as initially proposed in [35]. However, the constraint under consideration possesses a highly nonlinear nature, which prompts the authors to recast it in a form more conducive to optimization. Consequently, the optimization problem is reformulated by integrating the constraint within the objective, as shown below:

$$\min \|\delta\|_p + c * f(x + \delta). \quad (4)$$

In the context of the chosen distance metric $\|\cdot\|_p$ with L_p norm and a suitably selected constant parameter $c > 0$, while maintaining the second constraint from [35] unaltered, equation 4 yields more potent AEs when tackled with gradient descent in comparison to the FGSM.

The implementation nuances diverge based on the employed metric as follows:

- The L_2 norm implementation incorporates multiple initial points for the gradient descent to mitigate the chances of the algorithm landing in unfavorable local minima.
- As stated by Carlini and Wagner [18], the L_0 metric is nondifferentiable, necessitating the use of an iterative algorithm. This algorithm identifies the least important pixels in each iteration (utilizing the L_2 attack) and subsequently fixes them. Upon identifying a minimal subset of pixels, it is employed to generate an AE. This approach bears resemblance to JSMA, with the exception that while JSMA expands a set of alterable pixels, the C&W L_2 method reduces the pixel set.
- In line with Carlini and Wagner [18], the L_∞ metric lacks full differentiability and conventional gradient descent yields subpar results. This issue is circumvented with the use of an iterative

algorithm, replacing the $||\delta||_p$ term in the objective formulation with a penalty term $\tau = 1$. This revised objective is reevaluated at each iteration and provided $\delta < \tau$, the latter is reduced by a factor of 0.9, allowing transition to the subsequent iteration; otherwise, the search is terminated.

The full details of the implementation are complex and can be found in [18].

The experimental results presented by the authors demonstrate that each distance metric employed in the attacks yields AEs that are closer in comparison to those obtained from previous state-of-the-art attacks [18]. The L_0 and L_2 attacks produce AEs that exhibit $2\times$ to $10\times$ lower distortion compared to the best attacks reported in the literature, boasting a success probability of 100%. Although the L_∞ attacks yield AEs of similar quality to other studies, they outperform in terms of successful attack rates. Moreover, the effectiveness of the proposed techniques is such that their performance improves with increasing task complexity, a condition under which other methods typically deteriorate. The C&W attacks consistently achieve a 100% success rate on naturally trained DNNs across various datasets including MNIST, CIFAR-10, and ImageNet. Additionally, these attacks can successfully compromise defensive distilled models, a feat that DeepFool fails to accomplish in its search for adversarial samples [53].

2.5.4 PixelAttack

The One-Pixel Attack (PixelAttack) [19, 20] method, predicated on differential evolution, fabricates AE by perturbing a single pixel [19] or a limited number of pixels [20], utilizing either the L_0 or L_∞ metric. The attack delineated in these studies concentrates on a small number of pixels without restricting the intensity of modification. This attack aims to derive an AE x' from an original sample x via the computation of a minimal perturbation δ . This results in $x' = x + \delta$, $f(x) \neq f(x')$, where $f(\cdot)$ represents the model's output. Consequently, the goal of the associated optimization problem is [20]:

$$\min_{\delta} g(x + \delta)_c \text{ subject to } ||\delta|| \leq th, \quad (5)$$

where th denotes a prespecified threshold parameter that governs the maximum count of alterable pixels, while $g(\cdot)_c$ signifies the confidence associated with the correct class c , such that $f(x) = \arg \max g(x)$. The perturbations are encapsulated within arrays, referred to as candidate solutions, and are subjected to optimization through the process of differential evolution [19].

Each candidate solution comprises a constant number of perturbations where every perturbation alters a single pixel. This alteration is represented as a quintuple that includes the (x, y) coordinates and the RGB values associated with the perturbation. Upon generation, each candidate solution is pitted against its respective parent, based on the population index, and the victor persists into the subsequent iteration.

The potential evolution strategies include differential evolution and covariance matrix adaptation. As previously discussed, the attack methodology is designed around two distance metrics [20], which are detailed as follows:

- **Threshold Attack:** Leveraging the L_∞ metric, this attack can enact slight perturbations across

all pixels. It is constrained by the optimization of $\|\delta\|_\infty \leq th$, which allows the algorithm to search within the same space as the input, given that the variables can be any variation of the input, with a maximum limit of th .

- **Few-Pixel Attack:** Utilizing the L_0 metric, this attack can strongly perturb selected pixels. It is a variation of the original One-Pixel Attack [19], and it optimizes the constraint $\|\delta\|_0 \leq th$. In this scenario, the search space for the variables is reduced, as it is a combination of pixel values (dependent on channels 'c' in the image) and position (two values X, Y) for all of the th pixels.

The experimental findings reveal a considerable disparity in robustness when faced with L_0 and L_∞ norm attacks. The attack's effectiveness is remarkably high, even with exceedingly low threshold th values, requiring only a slight perturbation to successfully generate AEs [20].

2.5.5 BoundaryAttack

The Boundary Attack (BoundaryAttack) [21] is an iterative adversarial attack that operates without a gradient. It commences from a substantial adversarial perturbation and subsequently endeavors to minimize the L_2 distance between the original and perturbed examples, while maintaining the adversarial nature. Specifically, the attack begins with an image of the target class and alternates steps between moving the image along the decision boundary (maintaining its adversarial status) and steps moving towards the original image to discover incrementally smaller perturbations. The direction of the steepest ascent on the boundary surface, which is the direction where the model output alters most rapidly, is identified at each iteration by the attacker. The fundamental concept of the algorithm revolves around executing a rejection sampling with an appropriate proposal distribution P to identify incrementally smaller perturbations. During the k -th step, the aim is to draw a perturbation η^k from a maximum entropy distribution while adhering to the subsequent constraints:

1. The perturbed instance, denoted as \tilde{o} , is confined within the original input domain: $\tilde{o}_i^{k-1} + \eta_i^k \in \text{original domain}$;
2. The relative magnitude of the perturbation is represented by δ : $\|\delta^k\|_2 = \delta * d(o, \tilde{o}^{k-1})$;
3. The perturbation diminishes the Euclidean distance between the perturbed instance and the original input by a relative proportion ϵ : $d(o, \tilde{o}^{k-1}) - d(o, \tilde{o}^{k-1} + \eta^k) = \epsilon * d(o, \tilde{o}^{k-1})$.

Here, $d(o, \tilde{o}) = \|o - \tilde{o}\|_2^2$ signifies the Euclidean distance between the original and the perturbed instances, and o denotes the original instance.

The original formulation presents substantial complexity due to the challenges in sampling from the given distribution. In light of this, a more straightforward heuristic is employed as follows:

- Utilization of a Gaussian distribution $N(0, 1)$, which is independent and identically distributed;
- Perturbed samples undergo rescaling and clipping to ensure the satisfaction of constraints (1) and (2);

- The parameter η^k is projected onto a sphere centred at o , such that $d(o, \tilde{o}^{k-1} + \eta^k) = d(o, \tilde{o}^{k-1})$, thereby maintaining constraint (1). This is referred to as the orthogonal perturbation step;
- A modest progression is made towards the original image, ensuring that constraints (1) and (3) are upheld.

The algorithm, as elucidated earlier, pivots on two important parameters: the cumulative perturbation length δ and the step length ϵ in the direction of o . The intricacies of parameter modification are complex, hence, for a comprehensive understanding, the reader is directed to the original study [21].

Empirical evidence substantiates the efficiency of BoundaryAttack as a robust black-box attack. Its capability to locate AEs with minimal perturbations, which are challenging for human detection, underscores its potency. Moreover, its superiority over other gradient-based attacks is evidenced by its ability to bypass the computationally intensive task of gradient computation with respect to the input.

3 Results and Discussion

The ensuing discourse delineates the outcomes of the attacks under consideration, as previously introduced in Attack Algorithms. The ART library [54] provided the implementations for all the methodologies under scrutiny, facilitating the examination of a multitude of configuration parameters to comprehensively probe the potential of each technique. To maintain conciseness, the specifics of each attack’s results are relegated to the Supplementary Materials, where the experiments with diverse parameter values are documented. The primary focus of this section, however, remains the comparative analysis of the optimized attacks.

The evaluation metrics employed include the model’s accuracy over the manipulated samples, the average perturbation introduced to the AEs vis-a-vis the original samples, and the time required for processing. The first two metrics, also utilized in all ART library examples, serve as the primary measures for gauging the performance disparities between attacks and thus, were chosen as the principal metrics for assessing the efficacy of the methods implemented.

Accuracy serves to evaluate the performance of classification models. It is informally delineated as the ratio of correct predictions made by our model, specifically for multiclass classification. It is expressed as $\frac{\text{Number of Correctly Classified Samples}}{\text{Total Number of Samples}}$.

The mean perturbation introduced is computed as $\frac{1}{|T|} \sum_{s \in T} \frac{\sum_{f \in F_s} |adv^s f - sf|}{|F_s|}$. This denotes the average disparity between the original and altered features for each individual sample. Here, T symbolizes the test set, F_s signifies the feature set of sample s , adv^s is the AE derived from sample s , and sf indicates the value of feature f in sample s .

The training and evaluation of the models were conducted using an identical distribution of data for each split as mentioned in the preceding section, specifically 64/16/20% for training, validation, and testing, respectively. Detailed insights regarding the structure of the test for each dataset can be referred to in Section S4 of the Supplementary Materials. The performance of these models,

evaluated in terms of accuracy on the comprehensive test set, as well as on the subsets of male and female emotional speech, is presented in Table 3.

Initially, a comparative evaluation is conducted predicated on the metrics under consideration. Subsequently, a comprehensive analysis is carried out, factoring in the inherent properties of the diverse attack techniques.

In our data presentation, we focus solely on the attack configuration that results in the lowest accuracy parameter, indicative of optimal performance within an adversarial context. This approach is maintained even in light of performance variances across genders under different configurations. In instances where multiple configurations yield identical outcomes, our selection is guided by the configuration that produces the minimum average perturbation.

Table 4 encapsulates the optimal configurations for each dataset and attack, serving as a comparative reference in subsequent sections.

Attack	EmoDB	EMOVO	RAVDESS
FGSM	$eps = 1.25$	$eps = 0.5$	$eps = 0.25$
BIM	$eps = 0.25$	$eps = 0.25$	$eps = 0.25$
DeepFool	$iter = 5$	$iter = 5$	$iter = 5$
JSMA	$theta = +1$	$theta = +1$	$theta = +1$
C&W	$metric = L_{\infty}$	$metric = L_{\infty}$	$metric = L_{\infty}$
PixelAttack	$th = 10$	$th = 10$	$th = 10$
BoundaryAttack	-	-	-

Table 4: Best accuracy-performing configuration of each attack.

In general, the optimal configuration remains consistent across the entire test set, for both male and female samples. However, a few exceptions have been noted. For the EMOVO dataset, the FGSM attack outperforms others on the entire dataset and female samples when $eps = 0.5$. Yet, for male samples, equivalent accuracy is achieved when $eps = 0.25$, albeit with reduced perturbation. Similarly, the JSMA attack on the EMOVO dataset provides superior results on the entire dataset and female samples when $theta = +1$. However, male samples exhibit improved accuracy with $theta = -1$. Differently, for the RAVDESS dataset, the JSMA attack proves most effective on the entire dataset and male samples when $theta = +1$, while female samples show enhanced accuracy with $theta = -0.5$.

3.1 Performance Comparison

In this section, we evaluate the outcomes, taking into account each performance metric separately.

3.1.1 Accuracy

In the experiments, we computed the accuracy as $\frac{\text{Number of Correctly Classified Samples}}{\text{Total Number of Samples}}$. The precision of the models is assessed using the AEs derived from the instigated attacks, as detailed in Table 4. The resultant accuracy metrics are presented in Table 5 and illustrated in Figure 4.

Attack	Dataset	All	Female	Male
FGSM	EmoDB	0.109	0.121	0.089
BIM	EmoDB	0.067	0.065	0.070
DeepFool	EmoDB	0.061	0.061	0.070
JSMA	EmoDB	0.013	0.010	0.019
C&W	EmoDB	0.069	0.076	0.065
PixelAttack	EmoDB	0.547	0.603	0.454
BoundaryAttack	EmoDB	0.045	0.057	0.038
Original	EmoDB	0.909	0.918	0.895
FGSM	EMOVO	0.070	0.078	0.061
BIM	EMOVO	0.076	0.088	0.064
DeepFool	EMOVO	0.086	0.104	0.068
JSMA	EMOVO	0.022	0.024	0.020
C&W	EMOVO	0.085	0.068	0.102
PixelAttack	EMOVO	0.364	0.339	0.389
BoundaryAttack	EMOVO	0.076	0.070	0.082
Original	EMOVO	0.872	0.852	0.893
FGSM	RAVDESS	0.146	0.164	0.127
BIM	RAVDESS	0.059	0.060	0.057
DeepFool	RAVDESS	0.052	0.049	0.056
JSMA	RAVDESS	0.017	0.019	0.016
C&W	RAVDESS	0.060	0.058	0.062
PixelAttack	RAVDESS	0.322	0.419	0.225
BoundaryAttack	RAVDESS	0.204	0.202	0.205
Original	RAVDESS	0.911	0.911	0.911

Table 5: Accuracy obtained w.r.t all considered datasets with the original test data and with the best-performing configuration for each attack. The highest accuracies for each gender are shown in bold.

The analysis of the results reveals a substantial impact of all attack variants, including the rudimentary FGSM, on the models’ performance. Among these, PixelAttack exhibits a distinct efficacy, setting itself apart from the rest. Its effectiveness is particularly pronounced for EMOVO and RAVDESS, whereas it shows less impact on EmoDB. Despite this discrepancy, the minimal perturbations it introduces to the samples, as elaborated in the succeeding section, underscore its impressive performance. This idiosyncratic behavior may be attributed to the specific configurations employed, characterized by low thresholds denoted by th . Across all three datasets, the JSMA attack emerges as the most potent, reducing the accuracy to approximately 1-2%.

The performance of all attacks on the EmoDB dataset is reasonable. The simple yet effective FGSM reduces the accuracy to approximately 10%, while BIM, DeepFool, and C&W yield similar results at around 6.5%. Interestingly, BoundaryAttack, a black-box attack operating with less information, outperforms the latter, suggesting that a successful attack can be executed with minimal or no information about the target.

In the EMOVO dataset, FGSM exhibits an unusual efficiency, ranking second and surpassing its iterative BIM variant. The BIM-DeepFool-C&W trio performs similarly, achieving around 8%

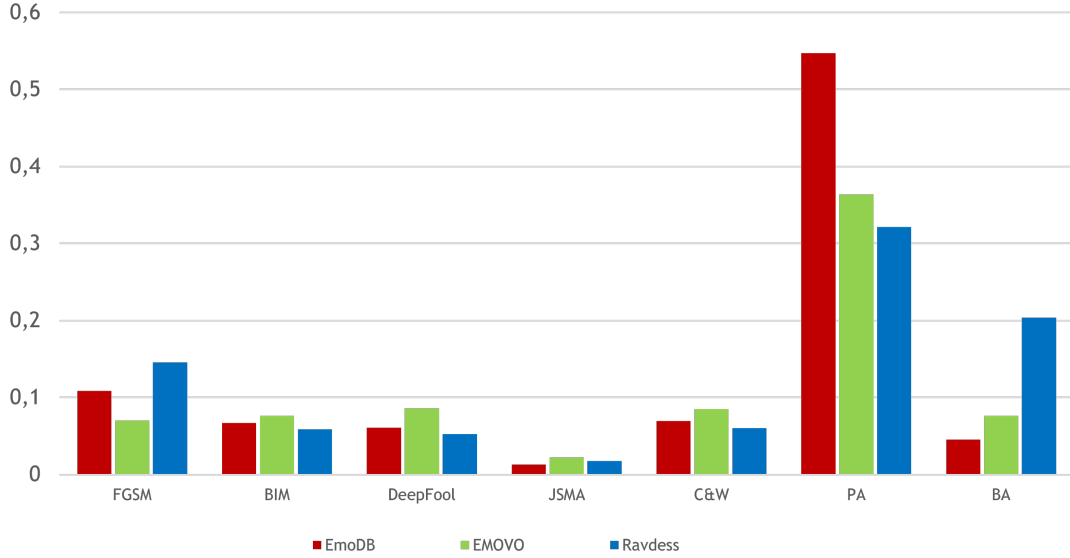


Figure 4: Accuracy obtained by the most effective configuration of each attack. Additional data can be found in Section S4 of the Supplementary Materials.

accuracy, on par with BoundaryAttack.

In the case of RAVDESS, BIM-DeepFool-C&W again generates similar results, this time around 5.6%, whereas BoundaryAttack does not maintain the same level of performance as observed in the other datasets.

In the comparative analysis of the three languages, the aggregate results exhibit remarkable similarity. The arithmetic mean of accuracies derived from white-box attacks by AEs is the lowest for EmoDB, registering at 0.063, followed by RAVDESS at 0.067, and lastly, EMOVO at 0.068. This allows us to infer with confidence that German, Italian, and English demonstrate equivalent susceptibility in the context of a SER task. Further exploration of this subject will be undertaken in the succeeding discourse.

Turning our attention to the variance between male and female samples, white-box attacks proved to be more successful on male subjects in 9 out of 15 instances. In the case of EMOVO, this trend is discernible in 4 out of 5 instances, a noteworthy observation given the higher accuracy for males compared to females in the original dataset. Likewise, for RAVDESS, male samples proved more susceptible in 3 instances. Conversely, for EmoDB, women were more impacted by 3 out of 5 attacks. The disparity is rather pronounced in certain attacks, with the accuracy differing by approximately 0.04 (excluding PixelAttack), while in other instances, the variation is around 0.01.

In summary, JSMA emerges as the most potent assault, substantially undermining the performance of models across all languages. The trio of BIM, DeepFool, and C&W exhibits a consistent performance across all cases. Even with its unassuming complexity, FGSM has demonstrated its efficacy across all three datasets, with a notable impact on EMOVO. Moreover, despite its black-box characteristics, BoundaryAttack achieves commendable results in two out of the three cases.

We have also evidenced how PixelAttack, by altering merely a handful of pixels, can induce large deviations in a model’s behavior.

The susceptibility of diverse languages to white-box attacks exhibits no marked disparities. The languages under scrutiny, namely German, Italian, and English, all demonstrate susceptibility to adversarial incursions. Furthermore, the analysis reveals that AEs derived from male audio samples typically yield higher efficacy, notably within the context of the EMOVO and RAVDESS datasets.

The augmentation of datasets was accomplished through the use of pitch shifting and time stretching techniques, inducing deformations to the input samples. Despite this, the deformations were not adequate to guarantee a robust defense against all forms of AEs attacks. These observations underscore the pronounced susceptibility of the SER task to such attacks when addressed using a CNN-LSTM model trained on log Mel-spectrograms.

3.1.2 Perturbation

We now turn our attention to the average perturbation induced by the attacks in the creation of the AEs. The data presented in Table 6 pertain to the highest accuracy outcomes derived from the parameter configurations encapsulated in Table 4 and depicted in Figure 5. For further inspection, we also provide samples of spectrograms perturbed in Figures 6-11.

Attack	Dataset	All	Female	Male
FGSM	EmoDB	1.070	1.070	1.068
BIM	EmoDB	0.159	0.159	0.158
DeepFool	EmoDB	1.894	1.917	1.858
JSMA	EmoDB	0.003	0.004	0.003
C&W	EmoDB	0.053	0.050	0.054
PixelAttack	EmoDB	5.55e-4	4.92e-4	6-6e-4
BoundaryAttack	EmoDB	0.757	0.775	0.746
FGSM	EMOVO	0.436	0.436	0.427
BIM	EMOVO	0.153	0.154	0.153
DeepFool	EMOVO	1.105	1.020	1.192
JSMA	EMOVO	0.002	0.002	0.002
C&W	EMOVO	0.035	0.035	0.035
PixelAttack	EMOVO	7.76e-4	8-03e-4	7.48e-4
FGSM	RAVDESS	0.198	0.204	0.192
BIM	RAVDESS	0.147	0.149	0.146
DeepFool	RAVDESS	1.327	1.411	1.243
JSMA	RAVDESS	0.002	0.003	0.002
C&W	RAVDESS	0.027	0.025	0.029
PixelAttack	RAVDESS	0.000939	0.000813	0.001066
BoundaryAttack	RAVDESS	1.314	1.493	1.136

Table 6: Mean perturbation injected by the best-performing configuration for each attack w.r.t all considered datasets. The entries in bold are the cases for which the best-performing configuration is also the less perturbing one.

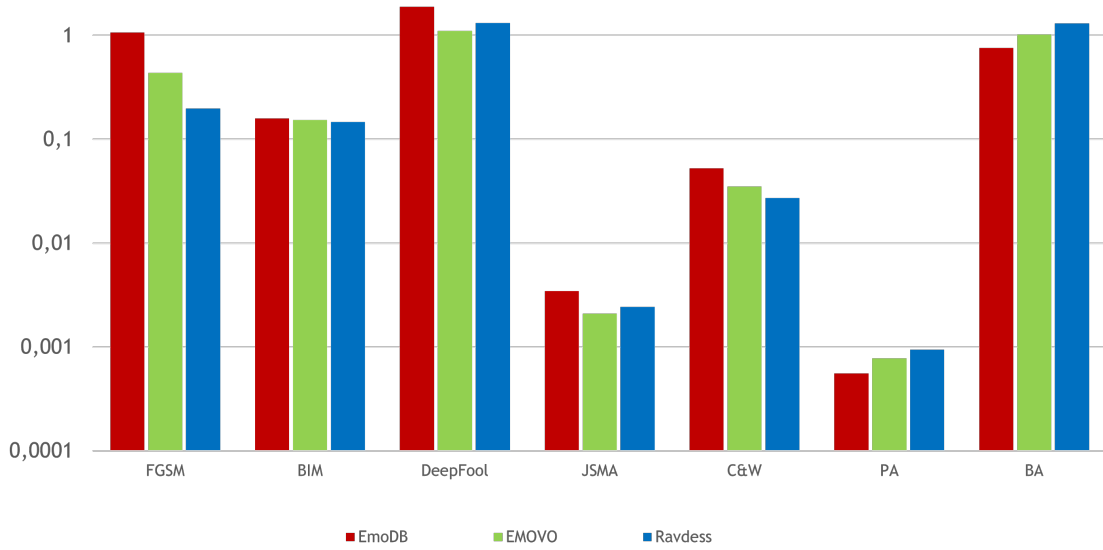


Figure 5: Perturbations injected by the most effective configuration of each attack. Additional data can be found in Section S4 of the Supplementary Materials.

The preceding section discussed the configuration of PixelAttack, which manipulates only a restricted set of pixels, thereby leading to a markedly diminished mean perturbation. Apart from PixelAttack, the most commendable performance is delivered by JSMA. Intriguingly, the configuration yielding the lowest accuracy for this attack is not the one introducing the minimal perturbation, but rather corresponds to $\theta = 0.5$. Further details are provided in Section S4 of the Supplementary Materials, which elucidates that each θ value corresponds to an average perturbation that is lower than that of all instances presented in Table 6. In the most unfavorable scenario, i.e., when $\theta = -1$, the outcomes are akin to those obtained with C&W using the L2 distance, albeit surpassing the results of the other attacks.

The triplet comprising BIM, DeepFool, and C&W, previously discussed for their comparable accuracy, demonstrates substantial discrepancies when it comes to the magnitude of introduced perturbations. Notably, an order of magnitude difference is observed between attacks for each dataset. In ascending order of perturbation magnitude, the attacks are C&W, BIM, and DeepFool.

Upon application to EmoDB, the optimal configuration of DeepFool, irrespective of gender, does not necessarily yield the least amount of noise introduced, albeit the discrepancy is negligible. In the case of RAVDESS, the least perturbation is attained with a solitary iteration of the algorithm, the difference being a mere 0.02. Comprehensive details for both instances are available in Section S4 of the Supplementary Materials. Remarkably, despite its nature as a black-box attack, BoundaryAttack exhibits performance on par with, and occasionally surpassing that of DeepFool.

The outcomes of the evaluated algorithms are generally consistent across various languages, with the exception of FGSM. This divergence can be attributed to performance fluctuations corresponding to different ϵ values, which dictate the attack step size. As anticipated, a decrease

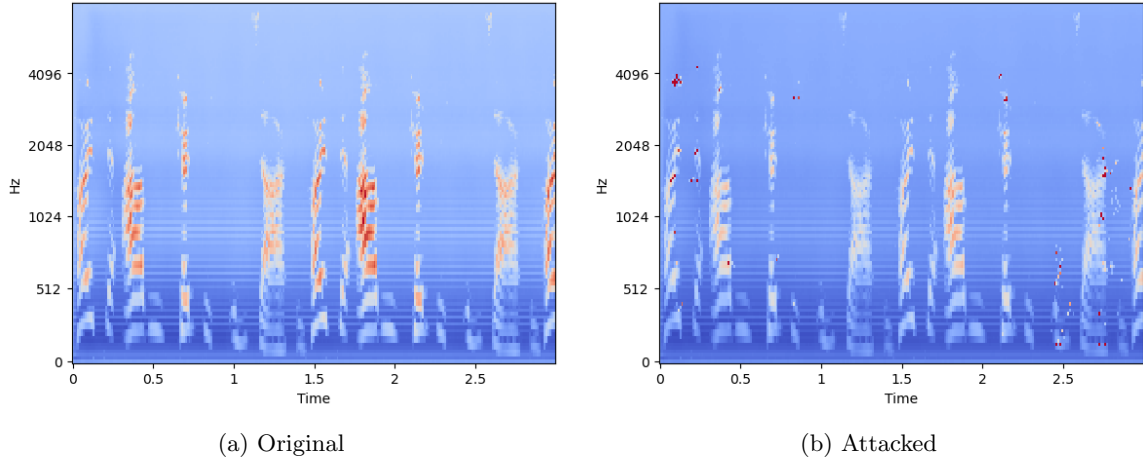


Figure 6: Example of male samples from EmoDB. (a) standardized original sample and (b) its JSMA-attacked version.

in *eps* values results in reduced perturbation levels. Based on the configurations outlined in Table 4, the perturbation levels span from optimal to suboptimal for RAVDESS, EMOVO, and EmoDB, respectively.

The influence of the speaker’s gender on all adversarial attacks is generally unimportant, with the notable exception of DeepFool, which exhibits the most substantial disparities.

To summarize, JSMA emerges as the most efficient attack mechanism, given its proficiency in inducing perturbations in the sample data. Its effectiveness is further underscored by the resultant decrease in accuracies, thereby solidifying its position as the preminent attack strategy.

As evidenced in Tables S23 and S24, DeepFool is characterized by the introduction of substantial noise. However, neither the accuracy nor the noise level exhibits noticeable enhancement with an escalation in the iteration count. The PixelAttack method, on the other hand, substantiates the feasibility of misleading the model through the alteration of a minimal number of pixels in the log Mel-spectrogram.

As Figures 6-11 exemplify, the resulting spectrograms are mainly degraded because of an almost uniform decrease of the amplitude of the spectrograms, resulting in sparse outlier points. When listening to such samples, the user is usually able to detect the attacked audio due to a noticeable decrease in volume. The GitHub repository contains examples of such audio.

3.1.3 Execution Time

In line with the evaluation carried out for precision and average disturbance, our attention now turns to the duration required by the under-consideration attack algorithms for the creation of AEs. The trial runs were conducted on a workstation of the HP Z4 G4 series, equipped with an i9-9820X CPU, a Nvidia TITAN V GPU possessing 12 GB of RAM, and a CPU RAM of 64 GB. The findings depicted in Figure 12 derive from the most effective precision outcomes garnered from the parameter setups delineated in Table 4.

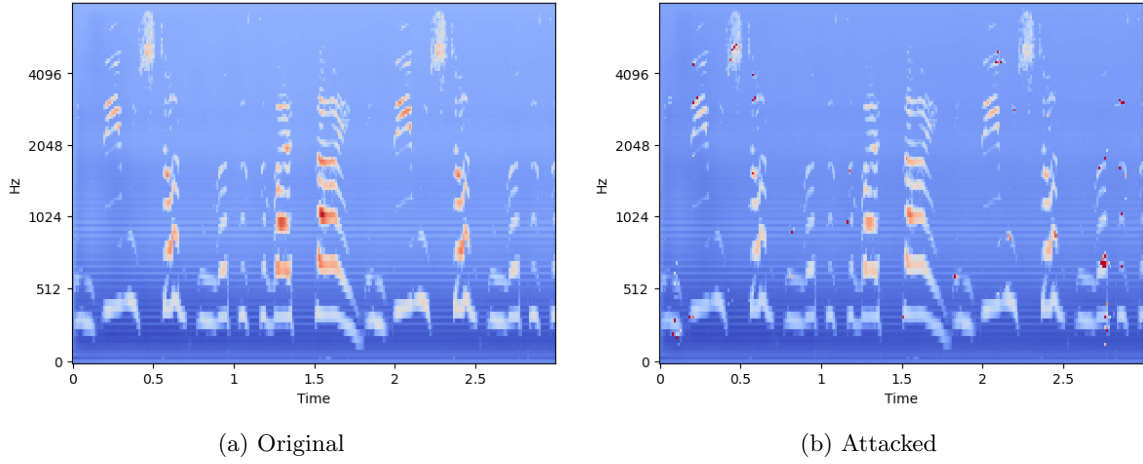


Figure 7: Example of female samples from EmoDB. (a) standardized original sample and (b) its JSMA-attacked version.

FGSM, as anticipated, outperforms all other attacks in terms of speed, even while introducing substantial noise. This makes it an effective and efficient approach for generating AEs rapidly. Although none of the selected configurations prove to be the fastest in terms of execution time, the disparities across various experiments are trivial, merely amounting to fractional seconds.

Focusing on the BIM-DeepFool-C&W trio, which we reiterate achieves comparable accuracy outcomes, noticeable variations are evident in terms of execution time aside from perturbation.

The C&W attack is the most time-consuming and requires a substantial duration to yield results. The scenario remains unchanged even when the L_2 distance is taken into account: the ensuing durations are akin to those of PixelAttack with $th = 1$, as highlighted in Section S4 of the Supplementary Materials.

Contrarily, DeepFool is capable of generating AEs promptly, although the perturbation induced is substantially high, as previously discussed.

BIM necessitates marginally extended durations compared to DeepFool, yet considerably less than C&W, positioning it as an optimal choice for creating high-quality AEs with minimal noise and in a reasonable timeframe. Although the configurations utilized are not the quickest, the disparities are inconsequential, akin to FGSM.

As previously noted, JSMA presents a high degree of effectiveness in impairing performance and introducing perturbation, although it requires a longer duration to generate samples in comparison to FGSM, BIM, and DeepFool. Nevertheless, its execution time remains considerably less than that of C&W. It is important to highlight that the parameter configurations selected to minimize accuracy concurrently result in reduced execution times. Despite the outcome for EMOVO not being strictly superior, the difference in execution time is a mere second.

Contrary to the majority of white-box attacks, the two black-box attacks necessitate a longer duration, except for C&W, which is the slowest technique overall. This is anticipated due to their limited insight into the internal workings of the target model.

For PixelAttack, it is pertinent to mention, as elaborated further in Section S4 of the Supple-

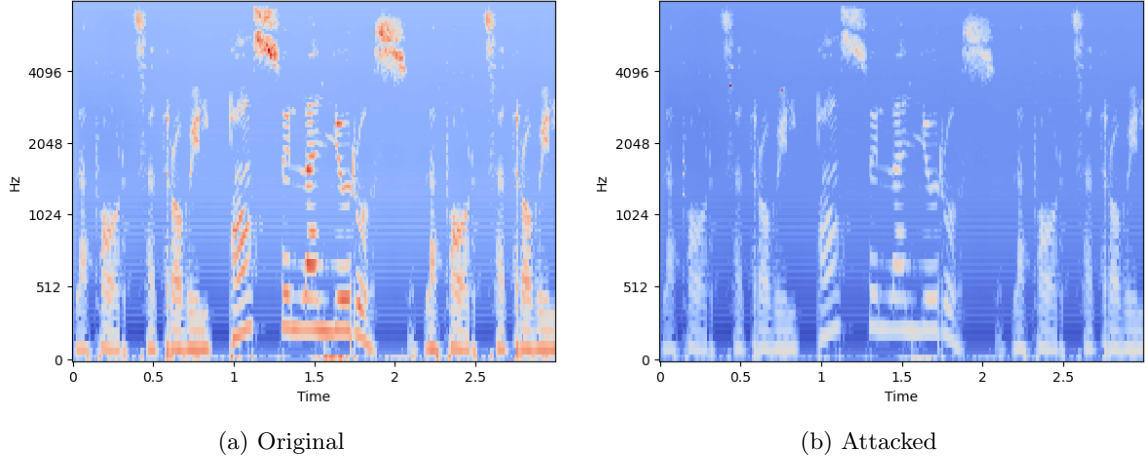


Figure 8: Example of male samples from EMOVO. (a) standardized original sample and (b) its JSMA-attacked version.

mentary Materials, that increasing the number of modifiable pixels reduces the computation time. As previously stated, it is important to consider that experimenting with larger values may lead to a decrease in both the model’s accuracy and the time required.

3.2 Comparison between Characteristics

In this section, a critical examination of the metrics under consideration is undertaken, with subsequent conclusions drawn from the intrinsic properties of each methodology and potential variances within the data.

3.2.1 Variations in Distance Metrics

It is imperative to understand that diverse attack types strive to minimize the disparity between the original samples, employing a range of distance metrics. A summary of the distance metrics utilized by the implemented ART is presented in Table 7.

Distance Metric	L_0	L_2	L_∞
FGSM			✓
BIM			✓
DeepFool		✓	
JSMA	✓		
C&W		✓	✓
PixelAttack	✓		
BoundaryAttack			✓

Table 7: Distance metrics used by the tested attacks.

Drawing definitive conclusions about potential disparities among distances and attacks utilizing the same distance metric is a complex task. Notably, JSMA and PixelAttack employ the L_0 distance,

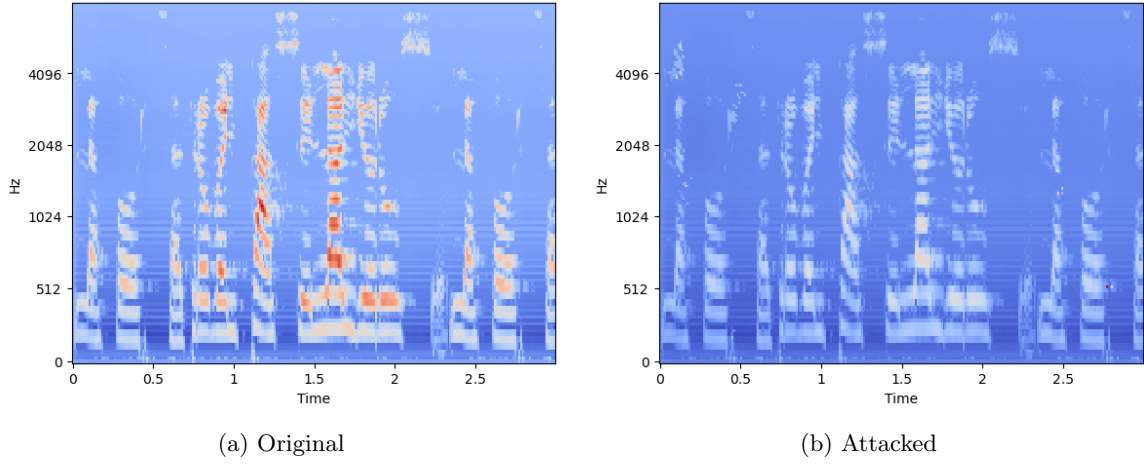


Figure 9: Example of female samples from EMOVO. (a) standardized original sample and (b) its JSMA-attacked version.

yet their effectiveness varies considerably. JSMA proves to be the most efficient, whereas PixelAttack is less effective. The gathered data indicates that these attacks introduce minimal perturbations, as they aim to reduce the quantity of altered pixels, a characteristic inherent to the L_0 distance.

Considering the L_2 category, it encompasses DeepFool and one variant of C&W. The interpretation of results becomes intricate here as the two attacks yield considerably divergent outcomes. Despite C&W achieving superior accuracy (tripling the score on EmoDB), the perturbation it introduces is remarkably lower (by three orders of magnitude). However, the generation of AEs via C&W is much more time-consuming, taking thousands of times longer than DeepFool. Consequently, it is challenging to extract consistent patterns from the executed experiments. The only conclusive remark is that the outcomes derived using the L_2 distance are profoundly influenced by the attack’s intrinsic logic.

Lastly, the L_∞ category, which encompasses the most substantial number of attacks, is considered. Despite all attacks striving to minimize the maximum discrepancy between the original and manipulated examples, the results exhibit considerable variations. The accuracy is comparable in all instances (with the exception of FGSM and BoundaryAttack on RAVDESS, which demonstrate notably superior results), yet the average perturbation and execution times differ considerably among various cases. Hence, the overall behavior of these techniques is primarily determined by their internal mechanisms.

The task of discerning a universal pattern through the juxtaposition of dissimilarities among attack groups utilizing identical metrics presents a considerable challenge. The primary source of variability in algorithms arises from the inherent methodology, rather than the minimized distance. The complexity is further heightened when attempting to compare attacks employing disparate metrics.

A preliminary inference drawn from the acquired results suggests that L_0 attacks appear to induce fewer perturbations compared to attacks that deploy other distance metrics. Yet, this inference warrants additional scrutiny, considering the fact that our testing was confined to merely two

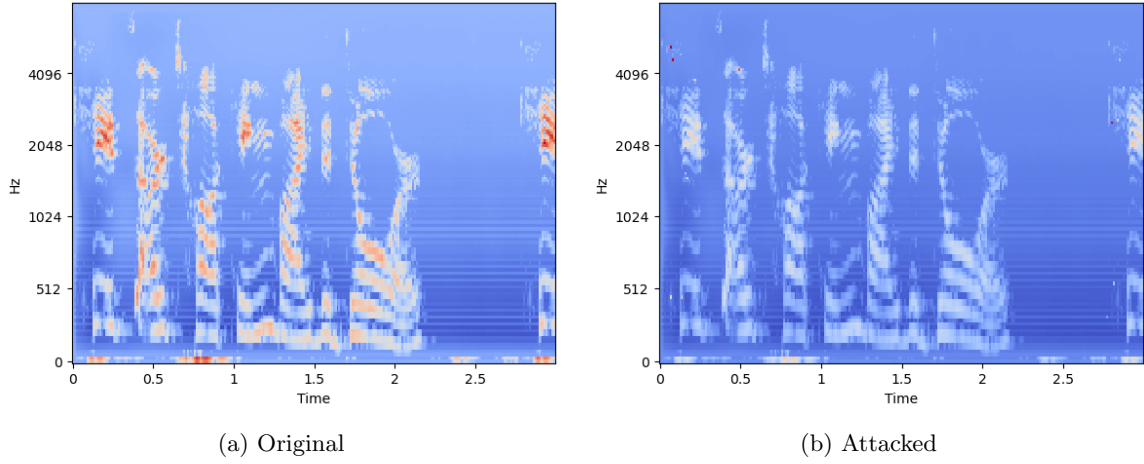


Figure 10: Example of male samples from RAVDESS. (a) standardized original sample and (b) its JSMA-attacked version.

algorithms within this category, and PixelAttack was set up differently compared to the rest of the techniques.

3.2.2 Iterative and Normal Versions

BIM and FGSM, two techniques with a comparative relationship, given that BIM is an extension of FGSM, exhibit differing performance under varying conditions. Upon examination of their accuracy results in light of all parameters considered, as presented in Tables S16 and S19 in Section S4 of the Supplementary Materials, it is notable that the *eps* parameter appears to exert no influence on BIM’s performance. This is evidenced by the consistent accuracy across different *eps* values. Conversely, FGSM’s accuracy fluctuates with the parameter, presenting a unique trend for each language. This suggests that BIM, with its independence from identifying the optimal configuration for deceiving the model, may be more advantageous.

The choice of datasets also influences the performance, as demonstrated in Table 5. BIM was observed to be more proficient in diminishing accuracy in the EmoDB and RAVDESS datasets, while FGSM demonstrated slightly superior performance in the EMOVO dataset.

With regard to the perturbation introduced, an increase in *eps* results in heightened noise for both attacks, as illustrated in Tables S17 and S21 in Section S4 of the Supplementary Materials. This is an anticipated outcome, given that the *eps* parameter signifies the maximum perturbation an attacker can introduce. Nevertheless, BIM generates more refined AEs with lower perturbation values.

Although BIM requires a much longer duration than FGSM to yield results, the attack time of 51 seconds for the largest dataset, RAVDESS, is deemed acceptable, as indicated in Tables S22 and S18 in Section S4 of the Supplementary Materials.

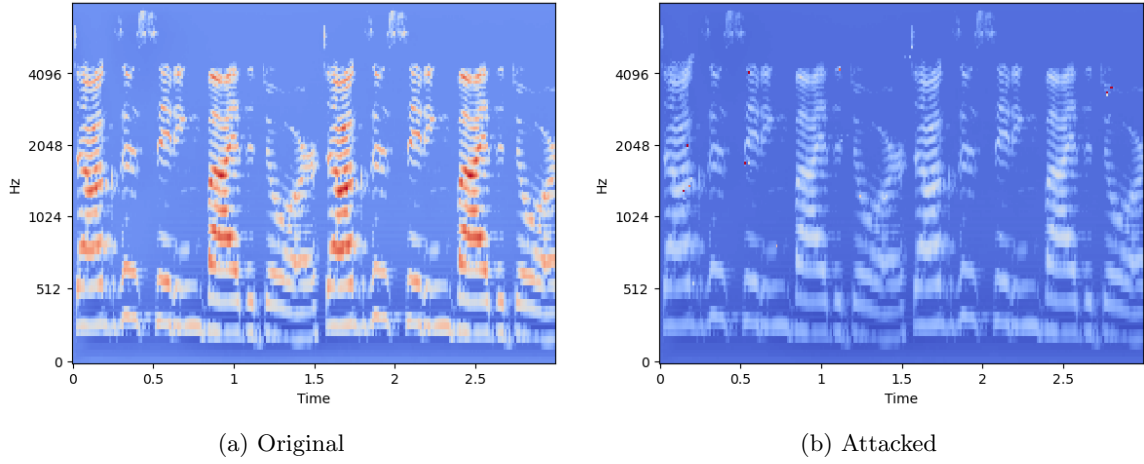


Figure 11: Example of female samples from RAVDESS. (a) standardized original sample and (b) its JSMA-attacked version.

3.2.3 White- and Black-Box Attacks

In this study, we executed a series of five white-box attacks and two black-box attacks. Despite the disparity in their quantities, a comparative analysis of these two categories is still important. This importance stems from the distinct configuration of PixelAttack, which leads to considerably dissimilar results compared to other forms of attack. Moreover, a closer inspection of the internal procedures utilized by these algorithms allows us to perceive this comparison in the context of gradient-based (white-box) and gradient-free (black-box) attacks, and the varying degrees of access to the model’s information.

Contrary to intuitive expectations, black-box attacks do not necessarily underperform due to their limited access to model information. This assertion is substantiated by the data presented in Table 5. For instance, the BoundaryAttack surpasses almost all white-box attacks in terms of performance for the EmoDB and EMOVO datasets. Nevertheless, this does not hold true for the RAVDESS dataset, where its performance is considerably inferior to all gradient-based attacks.

Upon evaluating the accuracy achieved by white-box attacks, it is evident that the results are comparable and consistently effective across different datasets. On the other hand, black-box techniques demonstrate wider variances. This indicates that gradient-based methods could potentially be language-independent, or at the very least, more so than population-based (PixelAttack) or decision-based (BoundaryAttack) methods, and are capable of working efficiently with log Mel-spectrograms.

From the average perturbation data presented in Table 6, it is evident that there are substantial variances in the performance of BoundaryAttack across different datasets. A similar trend is also discernible in the case of FGSM, where the selection of the *eps* value for accuracy minimization is of considerable importance, and DeepFool, where the noise level is even more pronounced. In contrast, PixelAttack consistently produces AEs with a comparable, low level of noise, given its configuration to alter only a minimal number of pixels.

On analyzing the perturbation variations across different languages, it is noted that black-box

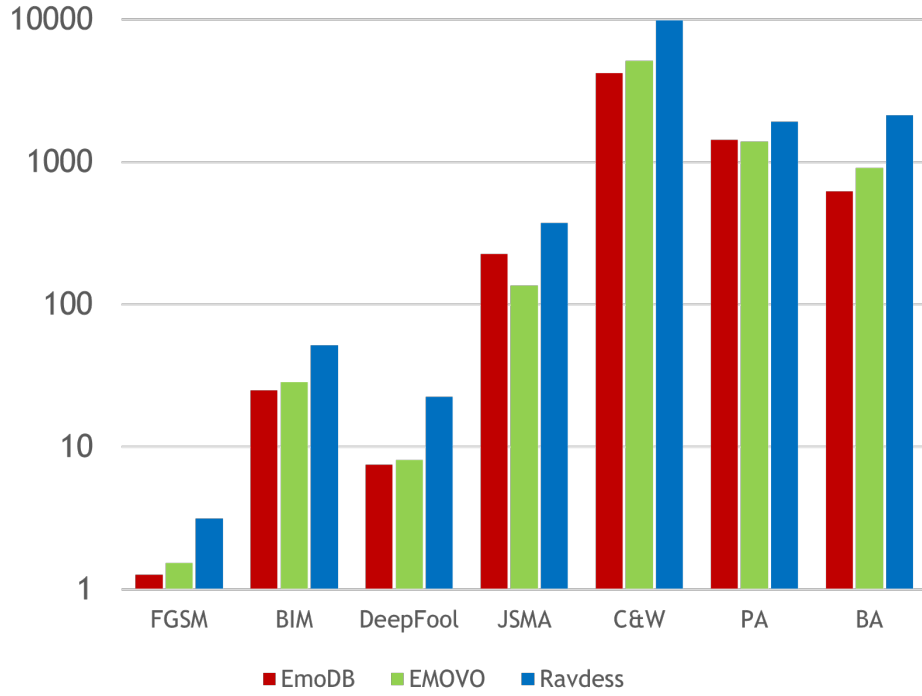


Figure 12: Time (s) required to generate the AEs for the various attacks and datasets for the best-performing configuration. Additional data can be found in Section S4 of the Supplementary Materials.

attacks introduce a lesser degree of perturbation for EmoDB, while the perturbation is more pronounced for white-box attacks. A more detailed discussion on this observation will be presented in the subsequent section.

As indicated in Table S37, the execution time for black-box attacks is typically longer, owing to the limited information available about the targeted model.

To summarize, despite the paucity of information about the victim model, black-box attacks can sometimes outperform their white-box counterparts by generating AEs with superior performance and lower disruption. However, our experimental results suggest that, on average, black-box attacks necessitate a longer execution time.

3.2.4 Differences between Languages

The trained models exhibit proficient performance across the three languages under consideration, as demonstrated in Table 3. RAVDESS yields the highest accuracy on the original data, registering at 0.912. This is closely followed by EmoDB and EMOVO, with respective accuracies of 0.909 and 0.872.

In Accuracy, we posited that the vulnerability to attacks across all three languages is relatively uniform. To elucidate this further, Figure 4 presents the accuracy across the three datasets for the most effective attack configurations, as detailed in Table S38, applied to the entire test set.

While the discrepancies between the achieved values are generally insubstantial, it is noteworthy that black-box attacks exhibit more important variations. Nevertheless, these observations enable us to derive some intriguing insights.

The model trained on the EMOVO dataset, despite exhibiting the lowest accuracy on the original data, outperforms the other models in terms of resistance to AEs. Specifically, it achieves superior performance when subjected to 4 out of 7 attack methods, thus suggesting a diminished impact of the attacks on this model. Consequently, it can be deduced that the model trained on Italian samples exhibits a marginally higher resilience.

In contrast, models trained on the RAVDESS dataset present a different scenario. These models, while attaining the highest accuracy on the original data, demonstrate a drop in performance when exposed to AEs, thereby making them the least resistant in 4 out of the 7 cases. This is particularly alarming given that the RAVDESS model was trained with a larger dataset and over a greater number of epochs, factors that would typically contribute to increased robustness. Crucially, this underscores the fact that the resilience of a model to AEs is not solely contingent on the volume of the training data, but also its quality.

An analogous analysis can be conducted on the injected perturbations. The accuracy of the most effective attack configurations across the three datasets, as presented in Table S39, is depicted in Figure 5 for the entire test set.

The EmoDB dataset is subjected to the most substantial perturbations in five out of all the attacks, implying that the majority of the implemented attacks have introduced the maximum level of noise. Notwithstanding the elevated perturbation, the attacks executed on EmoDB do not necessarily yield the lowest accuracies among the languages, as corroborated by Table S38. This observation suggests that the scrutinized techniques instigate an increased level of noise, which does not unequivocally translate into superior-performing AEs.

As previously alluded to, EmoDB exhibits a higher degree of perturbation across all white-box attacks, yet it manifests less interference under black-box attacks in contrast to the other two languages. This could imply that the efficacy of gradient-based methodologies in generating adversarial examples might be diminished when applied to the German language.

In terms of average perturbation, both EMOVO and RAVDESS demonstrate analogous scores, with the latter predominantly impacted by the outcomes of black-box attacks. By synthesizing the data from Tables S38 and S39 in Section S4 of the Supplementary Materials, it can be inferred that the assaults on RAVDESS are both effective (evidenced by low accuracy) and efficient (indicated by minimal noise introduction). This suggests that the English language might be more susceptible to AEs, and reinforces the notion that a model’s robustness does not necessarily equate to its resistance against such attacks.

EMOVO, on the other hand, registers the lowest average perturbation and, as anticipated, the highest accuracy score.

In summary, the present analysis demonstrates that AEs based on log Mel-spectrogram, when fed to a CNN-LSTM, can considerably degrade the performance of a SER model, regardless of the language considered. Although the performance differences among languages are relatively small, the experiments provide valuable insights.

Our findings indicate a heightened susceptibility of the English language to the discussed attacks, evidenced by its diminished accuracy notwithstanding its superior performance on the pristine data. Moreover, the observed mean perturbation is relatively unimportant, implying the generation of high-quality AEs.

Conversely, the Italian language demonstrates a greater degree of resilience to the same attacks, as inferred from its marginally superior accuracy and diminished perturbation.

The German language, however, presents a scenario that lies intermediate to the aforesaid languages. It exhibits an increased vulnerability specifically to gradient-based attacks, given that the perturbation introduced in these instances surpasses that noted for the other languages.

3.2.5 Differences between Genders

Table 3 illustrates that the trained models distinguish between male and female samples with negligible variations in the EmoDB original data. However, a pronounced discrepancy is discernible in the EMOVO dataset. Conversely, the models exhibit small fluctuations in the classification of male and female samples in the RAVDESS dataset.

Although gender disparities are generally nuanced, Table 13 provides a more detailed insight by presenting the accuracy exclusively for male/female samples across each dataset. Additional details are elaborated in Section S4 of the Supplementary Materials.

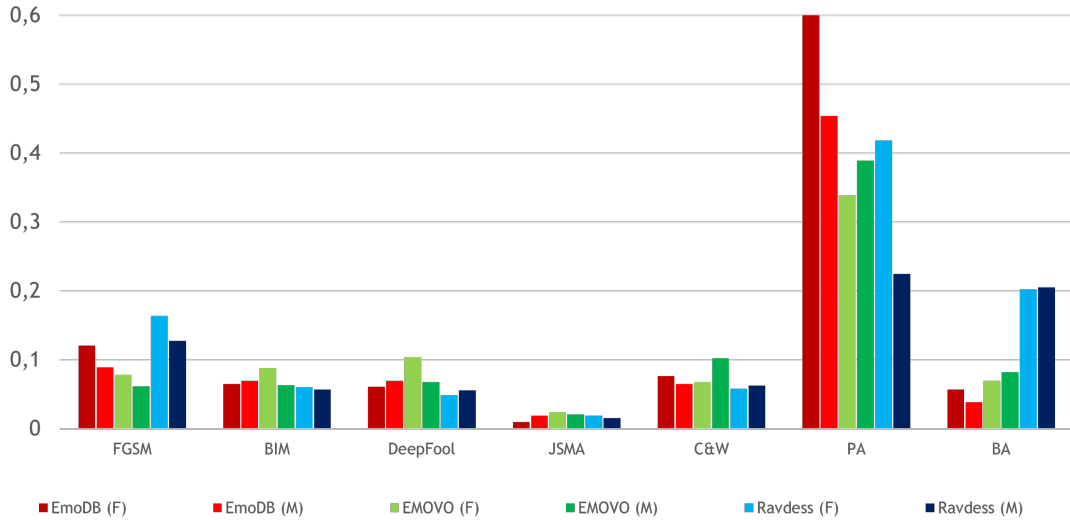


Figure 13: Accuracy obtained by the most effective configuration of each attack for male/female data across datasets.

A salient observation is that the utilization of PixelAttack yields large disparities between the two genders. This phenomenon could be attributed to the specific configuration employed. However, it implies that for marginal deviations from the original samples, both the German and English languages exhibit increased resilience towards female AEs, conversely, the Italian language demonstrates

greater resistance against male AEs.

Our preceding analyses, delineated in Accuracy, revealed a distinct pattern of white-box attacks exerting a greater impact on male subjects in 9 out of 15 instances – additional details can be found in Section S5 of the Supplementary Materials. This pattern was not exclusive to the aforementioned cases but was also observed in 4 out of 5 instances within the EMOVO dataset, despite the higher initial data accuracy of male subjects compared to their female counterparts. In the case of the RAVDESS dataset, the AEs proved to be more effective in 3 instances concerning male subjects, who, interestingly, exhibited marginally lower initial data accuracy. However, in the EmoDB dataset, a contrasting trend was observed. Here, 3 out of 5 attacks were more potent on female subjects, who had initially achieved higher accuracy scores on the original data compared to male subjects. In summary, the data suggests that the gender with superior initial data accuracy is more susceptible to attacks in two out of the three datasets analyzed. Consequently, this leads to diminished accuracy on AEs relative to the other gender. This indicates that the model’s resilience to gradient-based attacks on the best-performing gender cannot be reliably predicted solely based on the performance of the original data.

The outcomes of the black-box attack scenario present an equitable distribution, with males outperforming on the EmoDB, while females demonstrate superior attack efficacy on EMOVO. In the case of RAVDESS, each gender triumphs in one attack.

In addition, the findings corroborate those delineated in Table S38, which pertain to the most effective attacks on the datasets. This consistency in performance is observed even when the dataset is bifurcated into male and female categories. The minor variations in accuracy between the two genders do not considerably impact the overall efficacy of the attacks against the comprehensive AE dataset. Notably, RAVDESS comprises the majority of subsets with diminished accuracies, two attacks excel on EmoDB, while EMOVO records a single instance of superior performance with FGSM.

In a similar vein, Figure 14 provides perturbations exclusively for male and female samples within each dataset. Further elaboration on this topic is located in Section S4 of the Supplementary Materials.

An initial cursory examination suggests an insubstantial distinction between genders. Yet, a more meticulous analysis of the white-box results uncovers that in 11 out of 15 scenarios, male AEs manifest a diminished level of perturbation compared to female AEs. In addition, in 7 out of these 11 instances, males also demonstrate a lower accuracy rate than females, as illustrated in Table S40 in Section S4 of the Supplementary Materials. From these observations, it can be deduced that male AEs typically yield superior quality in terms of both the degradation of model performance and the magnitude of induced perturbation.

In the context of black-box attacks, males registered a lower accuracy in 4 out of 6 situations, although their accuracy was only inferior to females in a single case.

Upon evaluating individual attacks, it is discernible that male speech consistently manifests diminished perturbation across all instances for FGSM, BIM, JSMA, and BoundaryAttack. Conversely, for female speech, this phenomenon is solely observed with the C&W attack.

In summary, the findings suggest that there are negligible differences in performance between

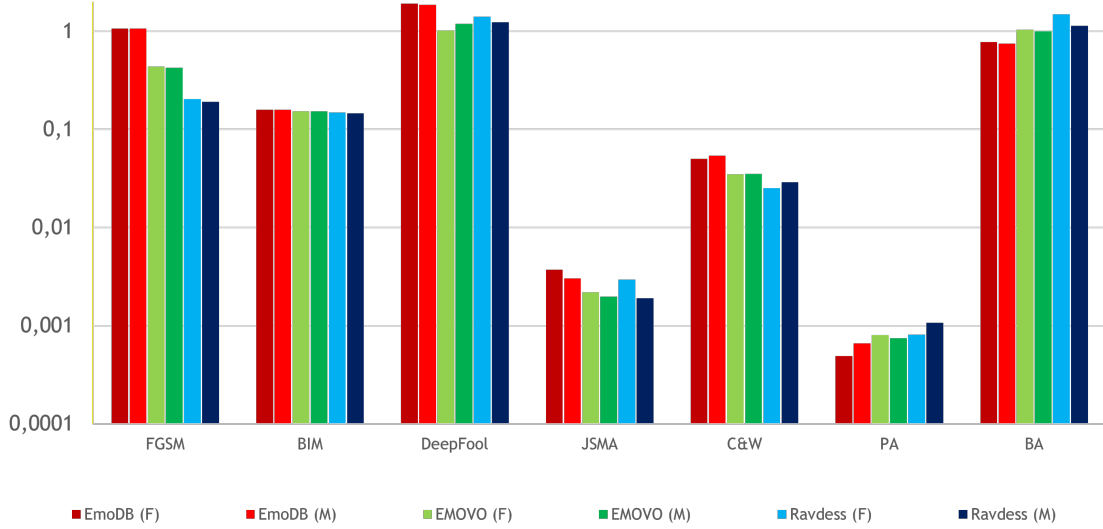


Figure 14: Perturbation injected by the most effective configuration of each attack for male/female data across datasets.

males and females in the majority of instances. Nevertheless, upon a more detailed examination, males seem to have a slight edge over females in terms of efficacy (lower accuracy) and quality (reduced induced noise), particularly in the context of white-box attacks.

4 Conclusion

The scientific community has been increasingly focusing on adversarial machine learning in recent years. Despite the surge in the development and application of new techniques for SER, the susceptibility of these methods to various forms of attack has not been sufficiently explored. This paper aims to fill this research gap by evaluating the robustness of SER systems against AEs. Our study scrutinizes three languages—German, Italian, and English—sourced from distinct datasets (EmoDB, EMOVO, and RAVDESS, respectively) to discern potential disparities among them. Furthermore, we have incorporated a gender-based perspective into our analysis, investigating the differential impacts of adversarial attacks on male and female speech. Importantly, the implementation of the presented experimental setup is publicly available at https://github.com/LIMUNIMI/thesis_adversarial_ml_audio ensuring full reproducibility of the achieved results.

We devised a pipeline to standardize the samples across the three languages and extract log Mel-spectrograms. Our methodology involved augmenting the datasets using pitch shifting and time stretching techniques, while maintaining a maximum sample duration of 3 seconds. Specifically, we generated eight distinct versions of the processed data, each differing in the data normalization method applied. The outcomes of our experiments were highly encouraging, demonstrating that the CNN-LSTM models performed optimally and consistently when standardized log Mel-spectrograms

were used, across all datasets.

To address the SER task, we established a uniform CNN-LSTM architecture across all datasets, thereby ensuring methodological consistency for attack comparisons. Through rigorous experimentation with diverse configurations of the neural network, optimal performance was achieved with a modestly sized CNN and 256 bidirectional LSTM units. Subsequent hyperparameter tuning further refined the performance for each dataset. This design strategy yielded high accuracy results on the EmoDB, EMOVO, and RAVDESS test sets, with respective accuracies of 90.92%, 89.52%, and 91.76%. These findings underscore the efficacy of employing a CNN-LSTM network trained on log Mel-spectrograms for the SER task while being in line with the state of the art.

Upon completion of the model development phase, we assessed the vulnerability of the resultant models to the previously mentioned attacks, under varying parameter configurations. Our empirical investigation revealed a substantial susceptibility of the SER task to AEs. Each examined attack method, including the relatively straightforward FGSM or PixelAttack, which was designed to alter a minimal number of features, successfully misled the network’s predictions. In light of these findings, it is evident that the CNN-LSTM model did not exhibit resilience against any of the employed attack techniques. Consequently, we advocate for the exploration of more robust models or alternative training data to enhance system robustness.

Our research findings indicate that amongst the multitude of attacks considered, JSMA surfaced as the most potent. The optimal configuration of its parameters led to a considerable drop in accuracy rates, resulting in 1.31 2.23 and 1.73 for EmoDB, EMOVO, and RAVDESS datasets, respectively. Furthermore, JSMA introduces only a minuscule degree of perturbation into the AEs. It is second only to PixelAttack, which is specifically tailored to alter a minimal number of pixels per spectrogram, in achieving the least perturbation.

The comparative analysis between white-box and black-box methodologies revealed that black-box techniques exhibit superior performance and minimal perturbation in two out of the three cases, specifically with BoundaryAttack. This is notwithstanding their limited access to information about the targeted model. Using BoundaryAttack, we recorded considerable drops in the accuracy of EmoDB, EMOVO, and RAVDESS to 4.54 7.6 and 20.38 respectively. These observations are alarming as they imply that attackers can potentially achieve remarkable results without any understanding of the model’s internal operation, simply by scrutinizing its output.

When we evaluated the impacts of the attacks across the three languages, no substantial difference in performance was observed. However, the results suggest that English appears to be the most susceptible, while Italian displays the highest resistance.

The comparative analysis between male and female samples revealed only negligible variations. A meticulous examination of the results, however, indicates a slight superiority of male samples, particularly in white-box attack scenarios, where they exhibited marginally lesser accuracy and perturbation.

To encapsulate, we introduced a reliable and efficacious approach for the training of a deep neural network for SER and corroborated it on three distinct languages. Our exploratory study on the model’s susceptibility to various adversarial attacks unveiled substantial vulnerabilities to all examined techniques, even revealing critical deficiencies in the face of black-box attacks. The

empirical trials showed that the proposed method does not exhibit considerable disparities in attack performance across different languages or gender samples, but only minor variances.

This work has advanced the field of SER research through the application of deep learning, offering understanding of the resilience of CNN-LSTM models and the influence of AEs on them. The findings presented herein establish a foundation for future studies focused on the creation of sturdier SER techniques, the development of more competent and impactful attacks, the investigation of potential defense mechanisms, the in-depth analysis of vocal variations across diverse languages and genders, and an overall enhanced understanding of the SER task.

Acknowledgments

We thank NVIDIA Corp. for the donation of two Titan V GPUs.

Author Contributions

N. Facchinetti designed, implemented, and conducted the experiments, analysed the experimental results, and prepared the original draft. F. Simonetta designed the experiments, analysed the experimental results, and reviewed and edited the paper. S. Ntalampiras conceived the idea, analysed the experimental results, and reviewed and edited the paper.

Competing interests

The authors declare that there is no conflict of interest regarding the publication of this article.

Data Availability

The execution of all the experiments depicted in this article can be found at https://github.com/LIMUNIMI/thesis_adversarial_ml_audio.

Supplementary Materials

- S1. Dataset Processing
- S2. Model Architectures
- S3. Improved Model Performance
- S4. Attack Deployment Results
- S5. Comparison Between Attacks
- Tables S1 to S41

References

1. Mantegazza I and Ntalampiras S. Italian speech emotion recognition. In: *2023 24th International Conference on Digital Signal Processing (DSP)*. 2023:1–5. DOI: 10.1109/DSP58604.2023.10167766.
2. Ntalampiras S. Speech emotion recognition via learning analogies. *Pattern Recognition Letters* 2021;144:21–6.
3. Ntalampiras S. Toward language-agnostic speech emotion recognition. *Journal of the Audio Engineering Society* 2020;68:7–13.
4. Ntalampiras S. A transfer learning framework for predicting the emotional content of generalized sound events. *The Journal of the Acoustical Society of America* 2017;141:1694–701.
5. Nicolini M and Ntalampiras S. Gender-aware speech emotion recognition in multiple languages. In: *Pattern recognition applications and methods*. Springer Nature Switzerland, 2024:111–23. DOI: 10.1007/978-3-031-54726-3_7. URL: http://dx.doi.org/10.1007/978-3-031-54726-3_7.
6. Ntalampiras S. Adversarial attacks against audio surveillance systems. In: *2022 30th European Signal Processing Conference (EUSIPCO)*. 2022:284–8. DOI: 10.23919/EUSIPCO55093.2022.9909635.
7. Ntalampiras S. Adversarial attacks against acoustic monitoring of industrial machines. *IEEE Internet of Things Journal* 2023;10:2832–9.
8. Meng H, Yan T, Yuan F, and Wei H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* 2019;7:125868–81.
9. Zhao J, Mao X, and Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control* 2019;47:312–23.
10. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B, et al. A database of German emotional speech. In: *Interspeech*. Vol. 5. 2005:1517–20. DOI: 10.21437/interspeech.2005-446.
11. Costantini G, Iaderola I, Paoloni A, and Todisco M. EMOVO Corpus: An Italian emotional speech database. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014:3501–4. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/591_Paper.pdf.
12. Livingstone SR and Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 2018;13:e0196391.
13. Akhtar Z and Dasgupta D. A brief survey of adversarial machine learning and defense strategies. Technical Report No. CS-19-002 2019.

14. Goodfellow IJ, Shlens J, and Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint 2014.
<https://doi.org/10.48550/arXiv.1412.6572>.
15. Kurakin A, Goodfellow IJ, and Bengio S. Adversarial examples in the physical world. In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018:99–112.
16. Moosavi-Dezfooli SM, Fawzi A, and Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016:2574–82.
17. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, and Swami A. The limitations of deep learning in adversarial settings. In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2016:372–87.
18. Carlini N and Wagner D. Towards evaluating the robustness of neural networks. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017:39–57.
19. Su J, Vargas DV, and Sakurai K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 2019;23:828–41.
20. Kotyan S and Vargas DV. Adversarial robustness assessment: Why both L_0 and L_{∞} attacks are necessary. arXiv preprint 2019.
<https://doi.org/10.48550/arXiv.1906.06026>.
21. Brendel W, Rauber J, and Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint 2017.
<https://doi.org/10.48550/arXiv.1712.04248>.
22. Abbaschian BJ, Sierra-Sosa D, and Elmaghraby A. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 2021;21:1249.
23. Nakatsu R, Nicholson J, and Tosa N. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In: *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*. 1999:343–51. DOI: 10.1145/319463.319641.
24. Petrushin V. Emotion in speech: Recognition and application to call centers. In: *Proceedings of Artificial Neural Networks in Engineering*. Vol. 710. 1999:22.
25. France DJ, Shiavi RG, Silverman S, Silverman M, and Wilkes M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering* 2000;47:829–37.
26. Schuller B, Rigoll G, and Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE. 2004:I–577.

27. Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016:5200–4. DOI: 10.1109/ICASSP.2016.7472669.
28. Latif S, Rana R, Khalifa S, Jurdak R, and Epps J. Direct modelling of speech emotion from raw speech. arXiv preprint 2019.
<https://doi.org/10.48550/arXiv.1904.03833>.
29. Etienne C, Fidanza G, Petrovskii A, Devillers L, and Schmauch B. CNN + LSTM architecture for speech emotion recognition with data augmentation. arXiv preprint 2018.
<https://doi.org/10.48550/arXiv.1802.05630>.
30. Purwins H, Li B, Virtanen T, Schlüter J, Chang SY, and Sainath T. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing* 2019;13:206–19.
31. Pandey SK, Shekhawat HS, and Prasanna SM. Deep learning techniques for speech emotion recognition: A review. In: *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE. 2019:1–6. DOI: 10.1109/radioelek.2019.8733432.
32. Ren Z, Baird A, Han J, Zhang Z, and Schuller B. Generating and protecting against adversarial attacks for deep speech-based emotion recognition models. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020:7184–8. DOI: 10.1109/icassp40776.2020.9054087.
33. Chang Y, Laridi S, Ren Z, Palmer G, Schuller BW, and Fisichella M. Robust federated learning against adversarial attacks for speech emotion recognition. arXiv preprint 2022.
<https://doi.org/10.48550/arXiv.2203.04696>.
34. Osman I and Shehata MS. Few-shot learning network for out-of-distribution image classification. *IEEE Transactions on Artificial Intelligence* 2022:1–13.
35. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint 2013.
<https://doi.org/10.48550/arXiv.1312.6199>.
36. Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2013:387–402.
37. Gong Y and Poellabauer C. Crafting adversarial examples for speech paralinguistics applications. arXiv preprint 2017.
<https://doi.org/10.48550/arXiv.1711.03280>.
38. Latif S, Rana R, and Qadir J. Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness. arXiv preprint 2018.
<https://doi.org/10.48550/arXiv.1811.11402>.
39. Purwins H, Li B, Virtanen T, Schlüter J, Chang SY, and Sainath T. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing* 2019;13:206–19.

40. Taori R, Kamsetty A, Chu B, and Vemuri N. Targeted adversarial examples for black box audio systems. In: *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE. 2019:15–20.
41. Carlini N and Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text. In: *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE. 2018:1–7. DOI: 10.1109/spw.2018.00009.
42. librosa.stft. <https://librosa.org/doc/main/generated/librosa.stft.html>. Accessed: 2022-12-20.
43. Chatziagapi A, Paraskevopoulos G, Sgouropoulos D, et al. Data Augmentation Using GANs for Speech Emotion Recognition. In: *Interspeech*. 2019:171–5.
44. Sahu S, Gupta R, and Espy-Wilson C. On enhancing speech emotion recognition using generative adversarial networks. arXiv preprint 2018.
<https://doi.org/10.48550/arXiv.1806.06626>.
45. Latif S, Asim M, Rana R, Khalifa S, Jurdak R, and Schuller BW. Augmenting generative adversarial networks for speech emotion recognition. arXiv preprint 2020.
<https://doi.org/10.48550/arXiv.2005.08447>.
46. Salamon J and Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* 2017;24:279–83.
47. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015:1–9.
48. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint 2014.
<https://doi.org/10.48550/arXiv.1409.1556>.
49. Ioffe S and Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*. pmlr. 2015:448–56.
50. Tan JH, Hagiwara Y, Pang W, et al. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Computers in Biology and Medicine* 2018;94:19–26.
51. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, and Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research* 2017;18:6765–816.
52. Kurakin A, Goodfellow I, and Bengio S. Adversarial machine learning at scale. arXiv preprint 2016.
<https://doi.org/10.48550/arXiv.1611.01236>.
53. Ren K, Zheng T, Qin Z, and Liu X. Adversarial attacks and defenses in deep learning. *Engineering* 2020;6:346–60.
54. Nicolae MI, Sinn M, Tran MN, et al. Adversarial Robustness Toolbox v1.2.0. CoRR 2018;1807.01069.