UDC 004.93

DEEPFAKE AUDIO DETECTION USING YOLOV8 WITH MEL-SPECTROGRAM ANALYSIS: A CROSS-DATASET EVALUATION

Zbezhkhovska U. R. – PhD, Leading Researcher of Scientific and Methodical Department for Quality Assurance of Educational Activities and Higher Education, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine.

ABSTRACT

Context. The problem of detecting deepfake audio has become increasingly critical with the rapid advancement of voice synthesis technologies and their potential for misuse. Traditional audio processing methods face significant challenges in distinguishing sophisticated deepfakes, particularly when tested across different types of audio manipulations and datasets. The object of study is developing a deepfake audio detection model that leverages mel-spectrograms as input to computer vision techniques, focusing on improving cross-dataset generalization capabilities.

Objective. The goal of the work is to improve the generalization capabilities of deepfake audio detection models by employing mel-spectrograms and leveraging computer vision techniques. This is achieved by adapting YOLOv8, a state-of-the-art object detection model, for audio analysis and investigating the effectiveness of different mel-spectrogram representations across diverse datasets.

Method. A novel approach is proposed using YOLOv8 for deepfake audio detection through the analysis of two types of mel-spectrograms: traditional and concatenated representations formed from SincConv filters. The method transforms audio signals into visual representations that can be processed by computer vision algorithms, enabling the detection of subtle patterns indicative of synthetic speech. The proposed approach includes several key components: BCE loss optimization for binary classification, SGD with momentum (0.937) for efficient training, and comprehensive data augmentation techniques including random flips, translations, and HSV color augmentations. The SincConv filters cover a frequency range from 0 Hz to 8000 Hz, with a step size of approximately 533.33 Hz per filter, providing detailed frequency analysis capabilities. The effectiveness is evaluated using the EER metric across multiple datasets: ASVspoof 2021 LA (25,380 genuine and 121,461 spoofed utterances) for training, and ASVspoof 2021 DF, Fake-or-Real (111,000 real and 87,000 synthetic utterances), In-the-Wild (17.2 hours fake, 20.7 hours real), and WaveFake (117,985 fake files) datasets for testing cross-dataset generalization.

Results. The experiments demonstrate varying effectiveness of different mel-spectrogram representations across datasets. Concatenated mel-spectrograms showed superior performance on diverse, real-world datasets (In-the-Wild: 34.55% EER, Fake-or-Real: 35.3% EER), while simple mel-spectrograms performed better on more homogeneous datasets (ASVspoof DF: 28.99% EER, WaveFake: 34.55% EER). Feature map visualizations reveal that the model's attention patterns differ significantly between input types, with concatenated spectrograms showing more distributed focus across relevant regions for complex datasets. The training process, conducted over 50 epochs with a learning rate of 0.01 and warm-up strategy, demonstrated stable convergence and consistent performance across multiple runs.

Conclusions. The experimental results confirm the viability of using YOLOv8 for deepfake audio detection and demonstrate that the effectiveness of mel-spectrogram representations depends significantly on dataset characteristics. The findings suggest that input representation should be selected based on the specific properties of the target audio data, with concatenated spectrograms being more suitable for diverse, real-world scenarios and simple spectrograms for more controlled, homogeneous datasets. The study provides a foundation for future research in adaptive representation selection and model optimization for deepfake audio detection.

KEYWORDS: deepfake detection, YOLOv8, mel-spectrogram, generalization capabilities.

ABBREVIATIONS

CNN is a convolutional neural network;

YOLO is a You Only Look Once model;

LA is a logical access;

DF is a deepfake;

PAN is a path aggregation network;

FPN is a feature pyramid network;

SPP is a spatial pyramid pooling;

BCE is a binary cross-entropy loss;

SGD is a stochastic gradient descent;

EER is a equal error rate metric;

STFT is a short-time Fourier transform;

FAR is a false acceptance rate;

FRR is a false rejection rate;

FP is a false positive;

TN is a true negative;

FN is a false negative;

TP is a true positive;

TTS is a text-to-speech;

VC is a vocoder;

HSV is a Hue, Saturation, and Value.

NOMENCLATURE

f is a frequency of a function;

v_t is a velocity term accumulating past gradients for momentum;

μ is a momentum coefficient for gradient updates;

θ_t is a model parameter at training step t;

η is a learning rate parameter controlling step size;

$\nabla_\theta L\left(\theta_t, x^{(i)}, y^{(i)}\right)$ is a gradient of the loss function L with respect to the parameters θ, computed for a single training sample $x^{(i)}$, $y^{(i)}$;

m is a mel scale value;

y is a ground truth label (0 for real audio, 1 for deepfake);

$\hat{y}$ is a predicted probability of the sample being a deepfake;

© Zbezhkhovska U. R., 2025
DOI 10.15588/1607-3274-2025-1-14

X is a set of audio signals $\{x_1, x_2, ..., x_N\}$;

Y is a set of corresponding labels $\{y_1, y_2, ..., y_N\}$.

INTRODUCTION

Detecting deepfake audio has become increasingly critical as the technology to create synthetic and altered speech has evolved. Deepfake audio can convincingly imitate human voices, often with the intention to deceive or manipulate, posing significant risks in areas such as security, media integrity, and public trust. Traditional audio processing methods face challenges in distinguishing deepfakes, particularly when tested across different types of audio manipulations and datasets.

One promising approach to address these challenges involves converting audio signals into visual representations, such as mel-spectrograms, which capture the sound's time-frequency features [2]. By transforming audio into images, computer vision models, which are highly effective at image recognition tasks, can be applied to detect patterns indicative of deepfake audio. CNN and other computer vision architectures can then analyze these spectrograms to detect anomalies or characteristics that differentiate genuine audio from deepfake audio. This method provides a novel and powerful approach to improve the accuracy of deepfake audio detection.

The object of study is developing a deepfake audio detection model that leverages mel-spectrograms as input to computer vision techniques.

Building such models requires significant computational resources, as training a network on large datasets of audio data is time-intensive. A major challenge in deepfake detection is ensuring that the model generalizes well, meaning it performs effectively not only on the dataset it was trained on but also on entirely new and unseen datasets. Many models perform well within their training environment but falter when encountering novel types of deepfakes, generalizing a key objective for practical deployment.

The subject of study is using mel-spectrograms in combination with computer vision models to enhance deepfake audio detection, focusing on improving the model's generalization across diverse datasets.

Current approaches to deepfake audio detection face significant challenges in generalization. Many models perform well on their training datasets but struggle when encountering new types of deepfakes or audio from different sources [2–27]. This limitation is particularly problematic given the rapid evolution of deepfake technologies. Mel-spectrograms offer a promising solution by transforming audio data into a visual format that can be analyzed using advanced computer vision techniques. These techniques have shown remarkable speed and accuracy in various image recognition tasks, but their full potential in deepfake audio detection via mel-spectrograms remains to be explored [28–29]. By investigating this approach, there is an opportunity to address the critical challenge of cross-dataset generalization in deepfake audio detection, potentially leading to more robust and versatile detection systems.

The purpose of the work is to improve the generalization capabilities of deepfake audio detection models by employing mel-spectrograms and leveraging computer vision techniques. By training the model on one dataset and testing it on others, this study aims to develop a more robust detection system that is effective across different types of deepfake audio.

1 PROBLEM STATEMENT

Suppose we are given an audio dataset represented as a set of instances $<X, Y>$, where $X=\{x_1, x_2, ..., x_N\}$ is the set of audio signals, and $Y=\{y_1, y_2, ..., y_N\}$ represents the corresponding labels, where $y_i=1$ for real audio and $y_i=0$ for fake audio. For each audio signal x_i, we convert it into a mel-spectrogram representation $S(x_i)$, such that the problem of deepfake detection can be transformed into an image classification problem using the mel-spectrograms.

Given this set of mel-spectrograms $<S(X), Y>$, the task is to train a computer vision model $F(\theta, S(x_i))$, where θ represents the set of model parameters, to predict whether an audio sample is real or fake. The objective is to minimize a loss function $L(F(\theta, S(x_i)), y_i) \to$ opt, where opt represents the optimal performance in terms of classification accuracy.

In addition, the problem of cross-dataset generalization is of primary interest. Specifically, for a model trained on a dataset $<S(X_{train}), Y_{train}>$, we aim to evaluate its performance on a distinct test set $<S(X_{test}), Y_{test}>$, where $X_{test} \neq X_{train}$ and the distribution of deepfake techniques may differ. The challenge is ensuring that the trained model generalizes well across diverse datasets, maintaining high accuracy on unseen data and addressing the limitations of dataset-specific detection methods.

2 REVIEW OF THE LITERATURE

The rapid advancement of artificial intelligence has led to the proliferation of deepfake audio, posing significant challenges to audio authenticity and security. Deepfake audio detection methods can be broadly categorized into pipeline approaches and end-to-end detectors [2–5]. Pipeline approaches involve a two-step process of feature extraction and classification. At the same time, end-to-end detectors aim to learn the detection task in a single step using deep neural networks.

Feature extraction techniques are crucial in capturing discriminative characteristics present in fake audio. These include short-term and long-term spectral features, prosodic features, and features derived from deep learning [6–8]. Short-term spectral features like Short-Time Fourier Transform effectively detect abrupt changes in audio signals [9], while prosodic features help uncover irregularities in speech pitch, intonation, and rhythm [10–11].

Recent advancements have incorporated self-supervised learning models like Wav2Vec, Wav2Vec2 XLS-R, and Hubert into the feature extraction process [12–15]. These models learn discriminative features from

DOI 10.15588/1607-3274-2025-1-14

raw audio without explicit labeling, potentially enhancing detection efficacy. However, a key challenge is ensuring these features generalize well across different types of deepfake attacks and audio datasets.

Traditional classifiers such as Support Vector Machines, Gaussian Mixture Models, and Logistic Regression have been employed in deepfake audio detection [16–18]. While these methods offer simplicity and efficiency, they often struggle to capture the intricate patterns introduced by sophisticated deepfake audio generation techniques, limiting their effectiveness against evolving attacks.

Deep learning approaches have shown significant promise in detecting subtle manipulations within audio data. CNN, particularly Light CNN, has performed excellently in deepfake audio classification tasks [19]. Residual Networks (ResNet) and its variants have also achieved promising results [20–21]. However, these models often require large amounts of training data and may not generalize well to unseen attack types. More advanced architectures like Res2Net [22], RawNet2 [23], and Squeeze-and-Excitation Networks [24] have been proposed to capture finer-grained audio features. Graph Neural Networks, such as RawGAT-ST, have improved performance in detecting a broad spectrum of spoofing attacks [25]. While these models offer enhanced feature learning capabilities, they often come at the cost of increased computational complexity and reduced interpretability.

A critical challenge in deepfake audio detection is the model's ability to generalize across different datasets and attack types. In [26], the authors observed that while their SincNet-based model performed well on known attacks, it struggled with attacks significantly different from those in the training set. This highlights the importance of diverse training data and robust evaluation protocols to ensure models can detect a wide range of deepfake techniques.

Transformer-based models like Rawformer have demonstrated improved performance and generalization across different datasets. The SE-Rawformer demonstrated good generalization, performing well on both ASVspoof 2019 LA and ASVspoof 2021 LA datasets [27]. However, the rapid evolution of deepfake technologies means that models must continuously adapt to new attack vectors, posing ongoing challenges for generalization.

While most research has focused on audio-specific architectures, the potential application of YOLOv8 to deepfake audio detection via mel-spectrogram transformation presents an interesting avenue for exploration. YOLOv8's efficiency in processing images could potentially translate to fast analysis of mel-spectrograms, enabling real-time deepfake audio detection. Its localization capabilities could be leveraged to identify specific segments of audio that have been manipulated. However, adapting YOLO from image detection to audio analysis may present challenges in capturing temporal dependencies and audio-specific features.

In the realm of deepfake video detection, YOLO-based approaches have shown promising results. The authors in [28] proposed a YOLO-CRNN-based deepfake detection approach that combines YOLO-Face for face detection with EfficientNet-B5 and Bi-LSTM for spatial-temporal feature extraction, achieving 89.38% accuracy and outperforming state-of-the-art methods on the CelebDF-FaceForensics++ (c23) dataset. Similarly, in [29] developed a YOLO-based framework for detecting manipulated faces in videos, demonstrating good generalization across different datasets. These successes in video deepfake detection suggest potential for adapting YOLO-based approaches to the audio domain, although careful consideration of the unique challenges in audio processing will be necessary.

3 MATERIALS AND METHODS

This study employs YOLOv8 [30], a state-of-the-art object detection model, for the task of deepfake audio detection. YOLOv8, known for its efficiency and accuracy in image recognition tasks, has been adapted to process mel-spectrograms derived from audio signals. The YOLO family of models has been at the forefront of real-time object detection, and YOLOv8 represents the latest iteration with significant improvements in both speed and accuracy.

YOLOv8 introduces several key enhancements over its predecessors, utilizing a new backbone network, CSPDarknet53 [31], which employ a cross-stage partial network to better balance accuracy and computational cost. The backbone of YOLOv8 is divided into four sections, each containing a single convolution layer followed by a C2f module [32]. It also integrates a PAN and a FPN for feature fusion, along with SPP to increase the receptive field. These architectural improvements allow YOLOv8 to capture multi-scale features, which are crucial for detecting deepfake artifacts in mel-spectrograms. The overall architecture thus comprises the backbone for feature extraction, the neck for fusing those features, and a head that generates bounding boxes and class predictions.

For our deepfake audio detection task, we adapt the YOLOv8 model to use BCE loss as the primary loss function [33]. The BCE loss is defined as:

$$BCE\left(y,\hat{y}\right)=-\left[y\cdot\log\left(\hat{y}\right)+\left(1-y\right)\cdot\log\left(1-\hat{y}\right)\right], \quad (1)$$

where y is the ground truth label (0 for real audio, 1 for deepfake) and $\hat{y}$ is the predicted probability of the sample being a deepfake. The BCE loss function is chosen for our deepfake audio detection task due to its ability to handle binary classification problems effectively. It measures the difference between the predicted probability and the actual label, guiding the model towards more accurate predictions.

For optimization, we employ SGD [34], which updates the model parameters by following the gradient of the loss function. The parameter update rule for SGD is mathematically defined as:

$$\theta_{t+1}=\theta_t-\eta\nabla_\theta L\left(\theta_t,x^{(i)},y^{(i)}\right). \quad (2)$$

© Zbezhkhovska U. R., 2025
DOI 10.15588/1607-3274-2025-1-14

In SGD, the gradients are computed using individual training samples or small batches of samples, which results in faster updates and more frequent parameter adjustments compared to full-batch gradient descent.

The learning rate η plays a crucial role in determining how large each update step is. A smaller learning rate provides more stable but slower convergence, while a larger learning rate speeds up training but risks overshooting the optimal parameter values.

We also introduce momentum to accelerate convergence and avoid oscillations during training. The momentum update rule modifies SGD as:

$$v_{t+1} = \mu v_t + \eta \nabla_\theta L\left(\theta_t, x^{(i)}, y^{(i)}\right),$$
$$\theta_{t+1} = \theta_t - v_{t+1}. \tag{3}$$

To adapt YOLOv8 for deepfake audio detection, we modify the final layers to output binary classifications (real or fake) instead of multiple object classes. To use audio signals with YOLOv8, we employ a multi-step approach. First, we convert the audio signals into mel-spectrograms. We then organize these mel-spectrograms into appropriate directory structures for YOLO training and create annotation files in YOLO format, specifying each spectrogram's class (real or fake).

The mel-spectrogram transformation is a critical step in our methodology, converting audio data into a visual format that can be analyzed by computer vision techniques. Mel-spectrograms represent the short-term power spectrum of sound based on a nonlinear frequency scale that approximates the human auditory system's response. This transformation allows us to capture temporal and frequency information in a format our adapted YOLOv8 model can effectively process.

The process of creating a mel-spectrogram involves several steps. First, the audio signal is divided into short, overlapping frames. We compute the STFT for each frame, which gives us the magnitude spectrum. This spectrum is then mapped onto the mel scale using a filterbank. The mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The conversion from frequency f to mel scale m is given by the equation:

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{4}$$

This transformation emphasizes lower frequencies, which are more perceptually significant in human hearing, and compresses higher frequencies. The resulting mel-spectrogram provides a compact representation of the audio signal that captures important features for deepfake detection.

Fig. 1 shows an example of a mel-spectrogram generated from an audio sample. The x-axis represents time, the y-axis represents mel frequency bands, and the color intensity indicates the energy level in each time-frequency bin. This visual representation allows our YOLOv8 model to identify patterns and anomalies that may indicate deepfake audio.

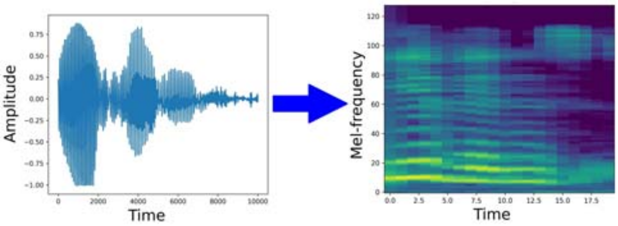

Figure 1 – Transformation of audio signal into mel-spectrogram

In addition to traditional mel-spectrograms, we also form mel-spectrograms as concatenated images from SincConv filters (Fig. 2) [35–36]. This approach allows us to leverage the benefits of learnable bandpass filters in the first layer of our neural network.

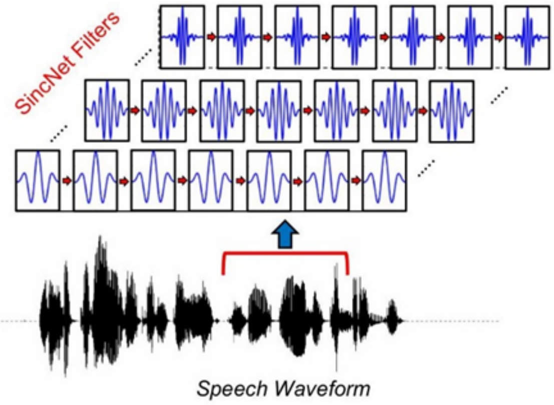

Figure 2 – SincConv filters

SincConv, or Sinc-based Convolutional Neural Networks, is a method introduced in [35] that uses sinc functions to implement band-pass filters in the first layer of a CNN. The SincConv layer learns the low and high cutoff frequencies of band-pass filters, which can be interpreted in the mel-scale, making it particularly suitable for our audio processing task.

The mathematical formulation of a SincConv filter is as follows:

$$h_{\sin c}(t) = \frac{\sin(2\pi f t)}{2\pi f}. \tag{5}$$

The SincConv layer applies these filters to the raw audio waveform, effectively learning to extract relevant frequency information. The output of this layer is then processed to form mel-spectrograms. Figure 3 illustrates concatenated mel-spectrograms formed by using SincConv filters. The image shows how 15 individual mel-spectrograms, each representing the output of a different SincConv filter, are combined into a single image. This

© Zbezhkhovska U. R., 2025
DOI 10.15588/1607-3274-2025-1-14

representation allows our YOLOv8 model to analyze multiple frequency bands simultaneously, potentially improving its ability to detect subtle artifacts in deepfake audio.

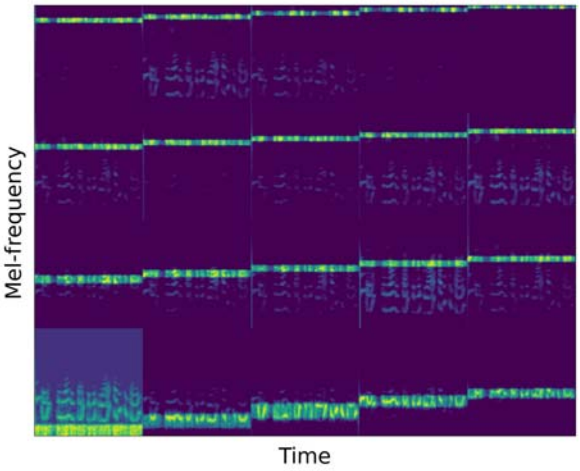

Figure 3 – Concatenated mel-spectrogram

By incorporating both traditional mel-spectrograms and those derived from SincConv filters, we provide our YOLOv8 model with rich, multi-dimensional representations of the audio signals. This approach aims to enhance the model's capacity to distinguish between genuine and deepfake audio by capturing a wider range of spectral and temporal features.

To evaluate the performance of our YOLOv8-based model, we use the EER metric [37]. EER provides a balanced measure of false positive and false negative errors, making it particularly suitable for assessing the effectiveness of deepfake detection models. The EER is calculated as the point where the false acceptance rate (FAR) equals the false rejection rate (FRR):

$$EER = FAR = FRR \tag{6}$$

where $FAR = \frac{FP}{(FP+TN)}$ and $FRR = \frac{FN}{(FN+TP)}$. FP represents False Positives, TN represents True Negatives, FN represents False Negatives, and TP represents True Positives. In practice, the EER is often determined by plotting the FAR and FRR curves and finding their intersection point. The lower the EER, the better the performance of the model.

4 EXPERIMENTS

Our methodology aims to address the critical challenge of cross-dataset generalization in deepfake audio detection. In this research we plan to train the YOLOv8 model on mel-spectrograms derived from one dataset and testing it on others, we seek to develop a more robust detection system that can effectively identify deepfakes across various audio sources and manipulation techniques. This approach has the potential to significantly enhance the practical applicability of deepfake audio detection in real-world scenarios.

To further explore the impact of different input representations on detection performance, we plan to train two YOLOv8 medium-size models. The first model will be trained using traditional mel-spectrograms, while the second will use concatenated mel-spectrograms formed from the output of SincConv filters. These SincConv filters cover a frequency range from 0 Hz to 8000 Hz, with a step size of approximately 533.33 Hz per filter. This comparison aims to assess whether the added frequency information provided by the SincConv-based mel-spectrograms enhances the model's ability to detect deepfakes across datasets.

Our study employs multiple datasets to ensure robust performance across various audio types and deepfake techniques. We primarily train our models on the ASVspoof 2021 LA dataset [6], which serves as a key benchmark in audio spoofing detection, introducing more advanced TTS and VC methods for synthetic speech generation. The LA partition contains 25,380 genuine and 121,461 spoofed utterances in the training set.

After training on the ASVspoof 2021 LA dataset, we assess model generalization by testing on the ASVspoof 2021 DF evaluation set and other datasets. One of these additional datasets is the "In-the-Wild" dataset [38], which contains fakes of politicians and public figures, sourced from publicly accessible platforms. This dataset includes 17.2 hours of fake audio clips and 20.7 hours of real audio clips. By incorporating real-world deepfakes, this dataset exposes the model to more diverse manipulation techniques, providing valuable insights into how well the model performs in uncontrolled environments.

We also test our models on the Fake-or-Real Dataset [39], which includes 111,000 real utterances sourced from open datasets, TED Talks, and YouTube, alongside 87,000 synthetic utterances generated by various TTS techniques. This dataset offers a broad variety of accents, recording conditions, and speech synthesis methods, enabling us to evaluate the model's performance under different scenarios. The diversity of real and fake utterances in this dataset further strengthens the evaluation by simulating a wide range of conditions that our model might encounter.

Additionally, we utilize the WaveFake dataset [40], which is composed entirely of synthetic speech generated by several TTS and VC architectures, including MelGAN, ParallelWaveGAN, HiFi-GAN, and WaveGlow. This dataset contains 117,985 fake audio files amounting to 196 hours of generated content. Though limited to a single speaker, WaveFake provides a focused evaluation of the model's ability to detect audio generated by modern speech synthesis techniques.

The training method parameters were set as follows: the number of epochs – 50, the loss function – BCE, and the learning rate – 0.01. SGD was employed as the optimizer, with a momentum of 0.937 and a weight decay of 0.0005. SGD was chosen for its simplicity and effectiveness in avoiding local minima, especially when combined with momentum, which accelerates convergence and helps the model navigate through flat regions of the cost

DOI 10.15588/1607-3274-2025-1-14

function. A warm-up learning rate strategy was applied, with a warm-up bias learning rate of 0.1 for the first 3 epochs to ensure smoother convergence.

To enhance generalization and model robustness, various data augmentation techniques were applied, inspired by methods like those in [41]. These included random flips, with a 50% probability of vertical flipping, and horizontal flipping disabled. Small translations and scaling were introduced, with values set to 0.1 for translation and 0.5 for scaling. HSV color augmentations were used, altering hue by 0.015, saturation by 0.7, and value by 0.4, reflecting potential variations in spectrogram images derived from different audio conditions. Additionally, the model applied mixup with a probability of 0, mosaic with a probability of 1.0, and random erasing with a probability of 0.4 to further diversify the training data. These techniques helped the model become more resilient to variations in audio spectrograms, improving its ability to generalize across different datasets.

After training, each model was evaluated on various test datasets, including the ASVspoof 2021 DF evaluation set, the "In-the-Wild" dataset, the Fake-or-Real dataset, and the WaveFake dataset. For each dataset, we computed the EER as the primary performance metric, which provided a balanced measure of false acceptance and false rejection rates.

To gain further insight into the model's behavior, we visualized feature extraction maps after each model layer. This visualization allowed us to observe how the model processed mel-spectrograms and concatenated representations from SincConv filters, providing deeper understanding into how it distinguished between real and fake audio signals during detection.

5 RESULTS

We present the results of our YOLOv8-based deepfake audio detection model, including visualizations of feature extraction maps and performance metrics.

Fig. 4 and Fig. 5 display the feature extraction maps from the most informative layers of the model when processing traditional mel-spectrograms and concatenated mel-spectrograms respectively. These visualizations highlight how the model captures essential low-level features and identifies key frequency patterns and temporal changes crucial for recognizing deepfake audio artifacts.

The performance of the YOLOv8 model was evaluated using the EER across different test datasets for both traditional and concatenated mel-spectrograms. The EER is a crucial metric for assessing the model's effectiveness in distinguishing between real and fake audio, providing a balanced measure of false acceptance and false rejection rates. The results are summarized in Table 1.

Table 1 – EER in % of YOLOv8 with different input mel-spectrograms

Model Dataset	YOLOv8 with simple mel-spectrograms	YOLOv8 with concatenated mel-spectrograms
ASVspoof DF	**28.99**	29.67
Fake or Real	39.58	**35.3**
In-the-wild	51.06	**34.55**
Wavefake	**34.55**	43.55

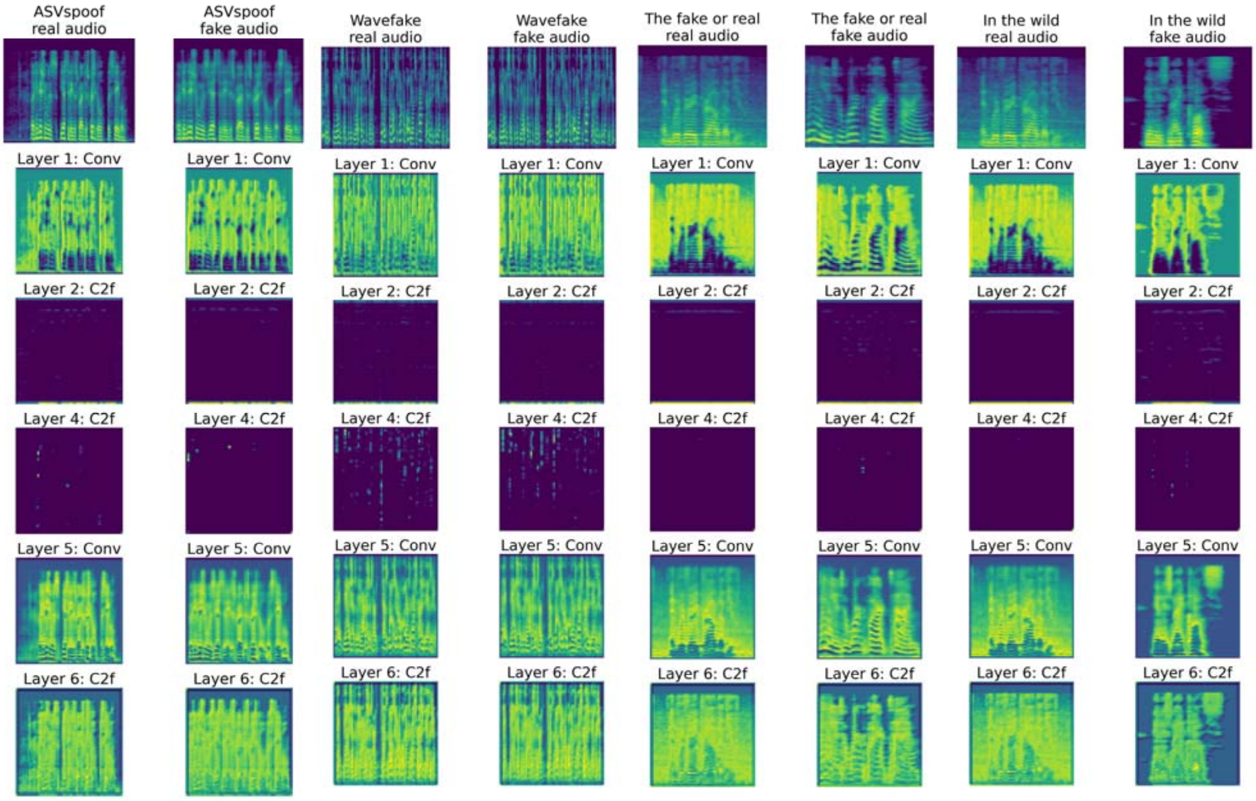

Figure 4 – The feature map of specific layers of YOLOv8 trained on simple mel-spectrograms

DOI 10.15588/1607-3274-2025-1-14

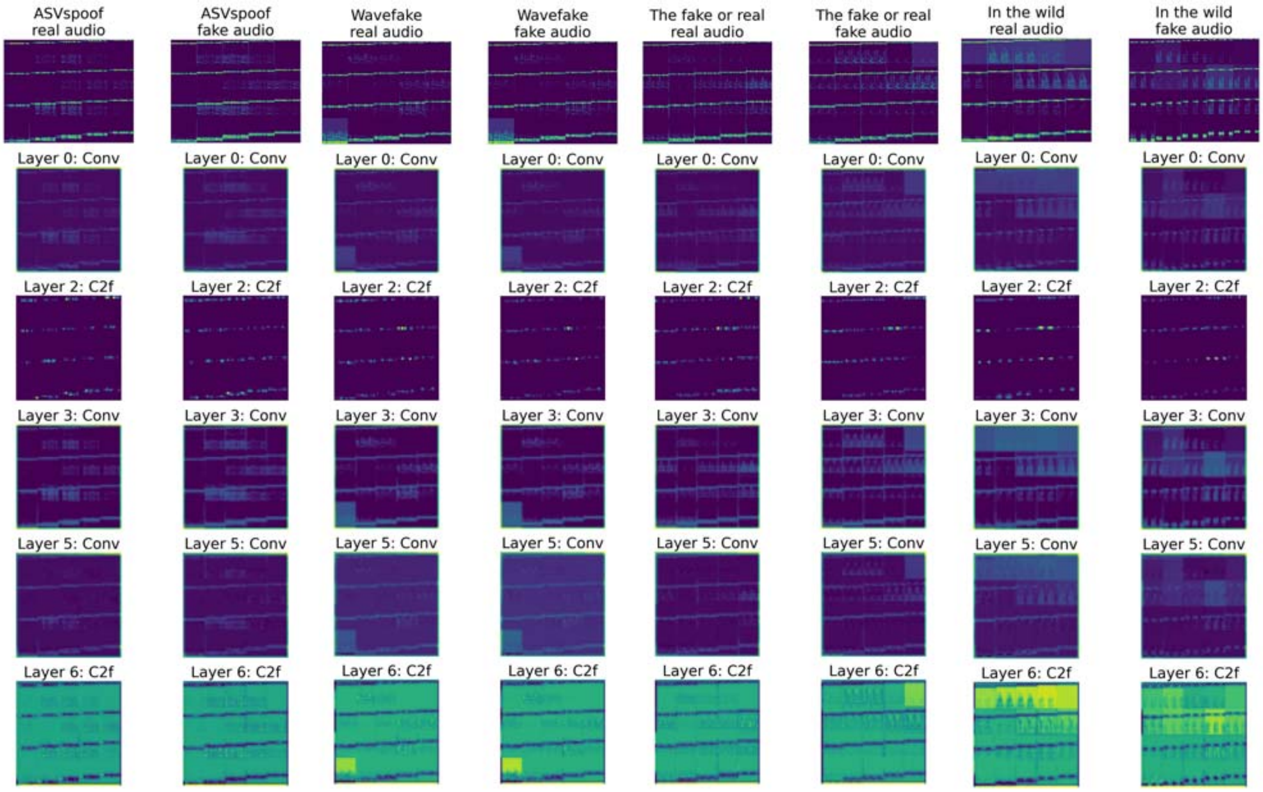

Figure 5 – The feature map of specific layers of YOLOv8 trained on concatenated mel-spectrograms

6 DISCUSSION

The results in Table 1, combined with the visualizations in Figures 4 and 5, offer valuable insights into how different input representations – simple versus concatenated mel-spectrograms – affect YOLOv8's ability to detect deepfake audio across various datasets. Comparing the results highlights the impact of input type on detection effectiveness, particularly when considering the unique characteristics of each dataset.

For the ASVspoof DF dataset, the EER with simple Mel-spectrograms was slightly lower (28.99%) than with concatenated spectrograms (29.67%). This minor difference suggests that the concatenated input does not add substantial value for ASVspoof DF, where simpler input captures most of the distinguishing features. Figures 4 and 5 support this observation; feature maps generated from both input types appear similar, indicating that YOLOv8 can recognize key deepfake patterns in ASVspoof DF equally well, regardless of input complexity. In this case, concatenated spectrograms do not provide any significant advantage, suggesting that simpler inputs may be sufficient for datasets with clear, identifiable deepfake characteristics.

In the Fake or Real dataset, however, concatenated spectrograms significantly improve performance, reducing the EER from 39.58% with simple spectrograms to 35.3%. This improvement likely stems from the model's ability to capture additional temporal and spectral information, as concatenation provides a richer context for identifying subtle deepfake cues. Figures 4 and 5 reflect this difference visually: the feature maps for concatenated spectrograms in Figure 5 show more detailed attention to distinctive regions, highlighting YOLOv8's enhanced ability to focus on nuanced patterns that simple spectrograms might overlook. This suggests that concatenated spectrograms are beneficial for datasets with higher variability, where added context helps distinguish genuine from fake samples.

The In-the-wild dataset shows the most substantial improvement with concatenated spectrograms, lowering the EER from 51.06% to 34.55%. This dataset is the most challenging due to its uncontrolled recording conditions and varied deepfake manipulations. Figures 4 and 5 illustrate how the model's focus is more effectively distributed across relevant regions when using concatenated spectrograms, which allows YOLOv8 to capture more complex, multi-dimensional features indicative of deepfake audio. In Figure 5, the feature maps reveal a more coherent and extensive focus across critical regions, demonstrating the model's improved capability to manage complex acoustic environments. This marked improvement with concatenated inputs underscores the importance of enhanced spectral-temporal representations when dealing with unpredictable, real-world data.

On the Wavefake dataset, in contrast, simple Mel-spectrograms produced a lower EER (34.55%) compared to concatenated spectrograms (43.55%), suggesting that the additional context from concatenated inputs may introduce noise rather than clarity. The homogeneity of the Wavefake dataset likely renders the additional informa-

DOI 10.15588/1607-3274-2025-1-14

tion unnecessary, and the model may perform best with a simpler, more focused input. Figures 4 and 5 further illustrate this difference, as feature maps in Figure 4 display a more targeted focus on specific regions for simple spectrograms, while the concatenated input in Figure 5 shows a diffused and less concentrated attention. This dispersion could explain the decrease in performance with concatenated spectrograms, as YOLOv8 may struggle to identify consistent patterns amidst additional, potentially irrelevant information.

Overall, the comparison of Figures 4 and 5 highlights the variability in model behavior across datasets with different input types. Concatenated Mel-spectrograms consistently offer improvements for datasets with greater variability (Fake or Real and In-the-wild), allowing YOLOv8 to capture intricate and temporally contextualized features that might otherwise go unnoticed. For datasets with more homogenous patterns, such as ASVspoof DF and Wavefake, simple Mel-spectrograms prove to be more effective by reducing noise and focusing the model's attention on specific, characteristic features. These results suggest that input representation choice should be tailored to dataset characteristics, with concatenated spectrograms being preferable for complex, varied data, while simpler spectrograms may suffice for more uniform datasets.

CONCLUSIONS

The urgent problem of deepfake audio detection is addressed through the development of a YOLOv8-based model that processes mel-spectrogram representations of audio signals.

The scientific novelty of obtained results is that a YOLOv8-based approach for deepfake audio detection is firstly proposed, which leverages both traditional and concatenated mel-spectrograms formed from SincConv filters. The model analyzes visual representations of audio signals to identify patterns indicative of synthetic speech. This approach demonstrates that computer vision techniques can be successfully adapted for audio authenticity verification, showing varying effectiveness across different types of datasets and mel-spectrogram representations.

The practical significance of obtained results is demonstrated through comprehensive experiments across multiple datasets, including ASVspoof 2021 DF, Fake or Real, In-the-wild, and Wavefake. The results reveal that the effectiveness of different mel-spectrogram representations varies significantly depending on the dataset characteristics. Concatenated mel-spectrograms showed superior performance on diverse real-world data (In-the-wild dataset, EER reduction from 51.06% to 34.55%) and the Fake or Real dataset (EER reduction from 39.58% to 35.3%). However, simple mel-spectrograms proved more effective for homogeneous datasets like Wavefake (34.55% vs 43.55% EER) and ASVspoof DF (28.99% vs 29.67% EER). This demonstrates the importance of selecting appropriate input representations based on the specific characteristics of the target audio data.

Prospects for further research include exploring additional mel-spectrogram formation techniques, investigating the impact of different YOLOv8 architectures, and developing adaptive methods that can automatically select the most appropriate mel-spectrogram representation based on dataset characteristics. Future work could also focus on improving the model's robustness against new types of audio deepfakes and reducing computational requirements while maintaining detection accuracy.

REFERENCES

1. Bondy M. Deepfakes, Digital Humans, and the Future of Entertainment in the Age of AI. *Perkins Coie.* URL: https://perkinscoie.com/insights/blog/deepfakes-digital-humans-and-future-entertainment-age-ai (date of access: 30.10.2024).
2. Mcuba M., Singh A., Ikuesan R. A., Venter H. The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation, *Procedia Computer Science*, 2023, № 219, P. 211–219. DOI: 10.1016/j.procs.2023.01.283
3. Salvi D. Liu H., Mandelli S., Bestagini P., Zhou W., Zhang W., Tubaro S. A Robust Approach to Multimodal Deepfake Detection, *Journal of Imaging*, 2023, Vol. 9, №6, P. 122. DOI: 10.3390/jimaging9060122
4. Wang C. Yi J., Tao J., Sun H., Chen X., Tian Z., Ma H., Fan C., Fu R. Fully Automated End-to-End Fake Audio Detection, *DDAM '22: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022, pp. 27–33. DOI: 10.1145/3552466.3556530
5. Wang X., Yamagishi J. Investigating Active-learning-based Training Data Selection for Speech Spoofing Countermeasure, *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 585–592. DOI: 10.1109/SLT54892.2023.10023350
6. Liu X., Wang X., Sahidullah M., Patino J., Delgado H., Kinnunen T. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, Vol. 31. pp. 2507–2522. DOI: 10.1109/taslp.2023.3285283.
7. Das R. K., Yang J., Li H. Long-range Acoustic and Deep Features Perspective on ASVspoof 2019, 2019 *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),* 2019, pp. 1018–1025. DOI: 10.1109/ASRU46091.2019.9003845
8. Wang X., Yamagishi J. Investigating Self-supervised Front Ends for Speech Spoofing Countermeasures, *The Speaker and Language Recognition Workshop*, 2021, pp. 100–106. DOI: 10.48550/arXiv.2111.07725
9. Xu L., Lee H., Chen Z., Chen X., Wang J. Modified Cepstral Feature for Speech Anti-spoofing, *Journal of Dong Hua University (English Edition)*, 2023, Vol. 40, № 2, pp. 193–201. DOI: 10.19884/j.1672-5220.202205007.
10. Leon P., Stewart B., Yamagishi J. Synthetic Speech Discrimination using Pitch Pattern Statistics Derived from Image Analysis, *Interspeech,* 2012, pp. 370–373. DOI: 10.21437/Interspeech.2012-135
11. Hong Y., Liu X., Tao J., Tian Z., Sun H. DNN Filter Bank Cepstral Coefficients for Spoofing Detection, *IEEE Access,* 2017, Vol. 5, pp. 4779–4787. DOI: 10.1109/ACCESS.2017.2687041
12. Schneider S., Baevski A., Collobert R., Auli M. Wav2Vec: Unsupervised Pre-training for Speech Recognition, *Interspeech*, 2019, pp. 1–9. DOI: 10.21437/interspeech.2019-1873

DOI 10.15588/1607-3274-2025-1-14

13. Babu A., Wang P., Mohamed A., Karthik M. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale, *Interspeech,* 2021, pp. 2278–2282. DOI: 10.21437/Interspeech.2022-143
14. Hsu W. N., Bolte B., Tsai Y., Salakhutdinov K., Mohamed T. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021. – Vol. 29. – P. 3451–3460.
15. Zbezhkhovska U., Khapilin O. Exploring Challenges and Future Paths in Deepfake Audio Detection, *Proceedings of the 5th Masters Symposium MS-AMLV-2024*. Lviv, Ukraine, March 29–30, 2024, pp. 1–10.
16. López J. A. V., Roddy M. P., Kinnunen T., Tan Z. Spoofing Detection with DNN and One-class SVM for the ASVspoof 2015 Challenge, *Interspeech,* 2015. DOI:10.21437/Interspeech.2015-468
17. Amin T. B., German J. S., Marziliano P. Detecting Voice Disguise from Speech Variability: Analysis of Three Glottal and Vocal Tract Measures, *Journal of the Acoustical Society of America,* 2013, Vol. 20. DOI: 10.1121/1.4879257
18. Rodríguez-Ortega Y., Ballesteros L. D. M., Renza D. A Machine Learning Model to Detect Fake Voice, *International Conference on Applied Informatics,* 2020, Vol. 1277, pp. 3–13. DOI: 10.1007/978-3-030-61702-8_1
19. Wu X., He R., Sun Z., Tan T. A Light CNN for Deep Face Representation with Noisy Labels, *IEEE Transactions on Information Forensics and Security*, 2018, Vol. 13, № 11, pp. 2884–2896. DOI: 10.1109/TIFS.2018.2833032
20. Alzantot M. F., Wang Z., Srivastava M. B. Deep Residual Neural Networks for Audio Spoofing Detection, *Interspeech*, 2019, pp. 1078–1082. DOI: 10.21437/Interspeech.2019-3174
21. Parasu P., Park K. K., Lee Y. C., Lee Y. Investigating Light-ResNet Architecture for Spoofing Detection under Mismatched Conditions, *Interspeech,* 2020, pp. 1111–1115. DOI: 10.21437/Interspeech.2020-2039
22. Gao S. H. Cheng M., Zhao K., Zhang X., Cheng M. M., Heng P. A. Res2Net: A New Multi-scale Backbone Architecture, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021, Vol. 43, №2, pp. 652–662. DOI: 10.1109/TPAMI.2019.2938758
23. Tak H., Patino J., Todisco M., Evans N. End-to-end Anti-spoofing with RawNet2, *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. – P. 6369–6373. DOI: 10.1109/ICASSP39728.2021.9414234
24. Lai C. I., Gao S. H., Chen Y. C. ASSERT: Antispoofing with Squeeze-Excitation and Residual Networks, *Interspeech*, 2019, pp. 1013–1017. DOI: 10.21437/Interspeech.2019-1794
25. Tak H., Patino J., Todisco M., Evans N. End-to-end Spectro-temporal Graph Attention Networks for Speaker Verification Anti-spoofing and Speech Deepfake Detection, *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 1–8. DOI: 10.21437/ASVSPOOF.2021-1.
26. Zeinali H., Mowlaee P., Patino J., Evans N. Detecting Spoofing Attacks using VGG and SincNet: Butomilia Submission to ASVspoof 2019 Challenge, *Interspeech*, 2019, pp. 1073–1077. DOI:10.21437/interspeech.2019-2892
27. Liu X., Yang J., Sun H., Wang L. Leveraging Positional-related Local-global Dependency for Synthetic Speech Detection, *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096278
28. Ismail A., Elpeltagy M. S., Zaki M. S., Eldahshan K. A. Deepfake Video Detection: YOLO-Face Convolution Recurrent Approach, *PeerJ Computer Science*, 2021, Vol. 7, pp. 2–19. DOI:10.7717/peerj-cs.730
29. Hubálovský Š., Trojovský P., Bacanin N., Venkatachalam Kv. Evaluation of Deepfake Detection using YOLO with Local Binary Pattern Histogram, *PeerJ Computer Science*, 2022. – Vol. 8. DOI: 10.7717/peerj-cs.1086.
30. Sohan M., Sai Ram T., Rami Reddy C. V. A Review on YOLOv8 and Its Advancements, *Data Intelligence and Cognitive Informatics,* 2024, pp. 529–545. DOI: 10.1007/978-981-99-7962-2_39.
31. Redmon J., Farhadi A. YOLOv3: An Incremental Improvement, *CoRR, arXiv preprint arXiv:1804.02767*, 2018. DOI: 10.48550/arXiv.1804.02767
32. Terven J., Córdova-Esparza D. M., Romero-González J. A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS, *Machine Learning and Knowledge Extraction,* 2023, Vol. 5, № 4, pp. 1680–1716. DOI: 10.3390/make5040083
33. Reis D., Kupec J., Hong J., Daoudi A. Real-Time Flying Object Detection with YOLOv8, *arXiv preprint arXiv:2305.09972*, 2024, pp. 1–10.
34. Ruder S. An Overview of Gradient Descent Optimization Algorithms, *arXiv preprint arXiv:1609.04747,* 2017, pp. 1–14.
35. Ravanelli M., Bengio Y. Speaker Recognition from Raw Waveform with SincNet, *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028. DOI: 10.1109/SLT.2018.8639585
36. Zbezhkhovska U. On Effectiveness and Generalization Capabilities of Deep Learning Models for Deepfake Audio Detection, *Master Thesis. Ukrainian Catholic University, Faculty of Applied Sciences, Department of Computer Sciences*. Lviv, 2024, 44 p.
37. Kinnunen T. H. Lee K. A., Tak H., Evans N. and Nautsch A. t-EER: Parameter-Free Tandem Evaluation of Countermeasures and Biometric Comparators, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, Vol. 46, № 5, pp. 2622–2637. DOI: 10.1109/TPAMI.2023.3313648
38. Müller N. M., Partel J., Kinnunen T. Does Audio Deepfake Detection Generalize?, Interspeech, 2022, pp. 2783–2787. DOI: 10.21437/Interspeech.2022-108
39. Reimao R., Tzerpos V. FoR: A Dataset for Synthetic Speech Detection, *2019 IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD),* 2019, pp. 1–10. DOI: 10.1109/SpeD.2019.8878970
40. Frank J., Schönherr L. WaveFake: A Data Set to Facilitate Audio DeepFake Detection, *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021, pp. 1–17. DOI: 10.5281/zenodo.5642694
41. Zhong Z., Zheng L., Kang G., Li S., Yang Y. Random Erasing Data Augmentation, *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020, pp. 13001–13008.

Received 18.11.2024.
Accepted 17.02.2025.

DOI 10.15588/1607-3274-2025-1-14

УДК 004.93

ВИЯВЛЕННЯ ГЛИБОКИХ ФЕЙКІВ В АУДІО ЗА ДОПОМОГОЮ YOLOV8 ТА МЕЛ-СПЕКТРОГРАМ

Збежховська У. Р. – д-р філософії, провідний науковий співробітник науково-методичного відділу забезпечення якості освітньої діяльності та вищої освіти, Харківський національний університет Повітряних сил імені Івана Кожедуба, Харків, Україна.

АНОТАЦІЯ

Актуальність. Проблема виявлення глибоких фейків у аудіо стає дедалі більш критичною в умовах швидкого розвитку технологій синтезу голосу та можливості їх використання з злочинною метою. Традиційні методи обробки аудіо стикаються з суттєвими викликами у виявлені складних аудіо фейків, особливо під час тестування на різних типах маніпуляцій з аудіо та наборах даних. Об'єктом дослідження є розробка моделі виявлення глибоких фейків у аудіо, яка використовує мел-спектрограми як вхідні дані для комп'ютерних методів зору, зосереджуючи увагу на покращенні можливостей узагальнення між наборами даних.

Мета роботи – покращення узагальнюючих можливостей моделей виявлення глибоких аудіо фейків шляхом використання мел-спектрограм та комп'ютерних методів зору. Це досягається шляхом адаптації YOLOv8, сучасної моделі комп'ютерного зору, для аналізу аудіо та дослідження ефективності різних представлень мел-спектрограм на різноманітних наборах даних.

Метод. Запропоновано новий підхід, що використовує YOLOv8 для виявлення глибоких аудіо фейків через аналіз двох типів мел-спектрограм: традиційних та конкатенованих, сформованих з фільтрів SincConv. Метод трансформує аудіосигнали в візуальні представлення, які можуть оброблятися алгоритмами комп'ютерного зору, що дозволяє виявляти тонкі шаблони, які свідчать про синтетичну мову. Запропонований підхід включає кілька ключових компонентів: оптимізацію функції втрат бінарної крос ентропії для задачі бінарної класифікації, стохастичний градієнтний спуск з моментом (0,937) для ефективного навчання та комплексні методи аугментації даних. Фільтри SincConv охоплюють частотний діапазон від 0 Гц до 8000 Гц з кроком приблизно 533,33 Гц на фільтр, забезпечуючи детальні можливості частотного аналізу. Ефективність оцінюється за допомогою метрики EER на кількох наборах даних: ASVspoof 2021 LA (25 380 справжніх та 121 461 підроблених висловлювань) для навчання, та ASVspoof 2021 DF, Fake-or-Real (111 000 реальних та 87 000 синтетичних висловлювань), In-the-Wild (17,2 години фейкових, 20,7 години реальних), та WaveFake (117 985 фейкових файлів) для тестування узагальнення між наборами даних.

Результати. Експерименти демонструють різну ефективність моделей в залежності від різних представлень вхідних даних. Конкатеновані мел-спектрограми продемонстрували кращу продуктивність на різноманітних реальних наборах даних (In-the-Wild: 34,55% EER, Fake-or-Real: 35,3% EER), тоді як прості мел-спектрограми працювали краще на більш однорідних наборах даних (ASVspoof DF: 28,99% EER, WaveFake: 34,55% EER). Візуалізації карт ознак показують, що шаблони уваги моделі значно різняться в залежності від типів вхідних даних, наприклад, конкатеновані мел-спектрограми демонструють більш розподілений фокус на відповідних областях для складних наборів даних.

Висновки. Експериментальні результати підтверджують доцільність використання YOLOv8 для виявлення глибоких аудіо фейків та демонструють, що ефективність представлень мел-спектрограм значно залежить від характеристик набору даних. Отримані результати свідчать, що представлення вхідних даних слід обирати на основі специфічних властивостей цільових аудіоданих, причому конкатеновані мел-спектрограми є більш підходящими для різноманітних реальних сценаріїв, а прості мел-спектрограми – для більш контрольованих однорідних наборів даних. Дослідження закладає основу для подальших досліджень у галузі адаптивного вибору представлення даних та оптимізації моделей для виявлення глибоких аудіо фейків.

КЛЮЧОВІ СЛОВА: виявлення глибоких фейків, YOLOv8, мел-спектрограми, узагальнюючі можливості.

ЛІТЕРАТУРА

1. Bondy M. Deepfakes, Digital Humans, and the Future of Entertainment in the Age of AI / M. Bondy // Perkins Coie. URL: https://perkinscoie.com/insights/blog/deepfakes-digital-humans-and-future-entertainment-age-ai (date of access: 30.10.2024).
2. The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation / [M. Mcuba, A. Singh, R. A. Ikuesan, H. Venter] // Procedia Computer Science. – 2023. – № 219. – P. 211–219. DOI: 10.1016/j.procs.2023.01.283
3. A Robust Approach to Multimodal Deepfake Detection / [D. Salvi, H. Liu, S. Mandelli et al.] // Journal of Imaging. – 2023. – Vol. 9, №6. – P. 122. DOI: 10.3390/jimaging9060122
4. Fully Automated End-to-End Fake Audio Detection / [C. Wang, J. Yi, J. Tao et al.] // DDAM '22: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. – 2022. – P. 27–33. DOI: 10.1145/3552466.3556530
5. Wang X. Investigating Active-learning-based Training Data Selection for Speech Spoofing Countermeasure / X. Wang, J. Yamagishi // 2022 IEEE Spoken Language Technology Workshop (SLT). – 2023. – P. 585–592. DOI: 10.1109/SLT54892.2023.10023350
6. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild / [X. Liu, X. Wang, M. Sahidullah et al.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2023. – Vol. 31. – P. 2507–2522. DOI: 10.1109/taslp.2023.3285283.
7. Das R. K. Long-range Acoustic and Deep Features Perspective on ASVspoof 2019 / R. K. Das, J. Yang, H. Li // 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). – 2019. – P. 1018–1025. DOI: 10.1109/ASRU46091.2019.9003845
8. Wang X. Investigating Self-supervised Front Ends for Speech Spoofing Countermeasures / X. Wang, J. Yamagishi // The Speaker and Language Recognition Workshop. – 2021. – P. 100–106. DOI: 10.48550/arXiv.2111.07725
9. Modified Cepstral Feature for Speech Anti-spoofing / [L. Xu, H. Lee, Z. Chen et al.], X. Chen, J. Wang // Journal of Dong Hua University (English Edition). – 2023. – Vol. 40, № 2. – P. 193–201. DOI: 10.19884/j.1672-5220.202205007.
10. Leon P. Synthetic Speech Discrimination using Pitch Pattern Statistics Derived from Image Analysis / P. Leon, B. Stewart, J. Yamagishi // Interspeech. – 2012. – P. 370–373. DOI: 10.21437/Interspeech.2012-135
11. DNN Filter Bank Cepstral Coefficients for Spoofing Detection / [Y. Hong, X. Liu, J. Tao et al.] // IEEE Access. – 2017. – Vol. 5. – P. 4779–4787. DOI: 10.1109/ACCESS.2017.2687041

DOI 10.15588/1607-3274-2025-1-14

12. Wav2Vec: Unsupervised Pre-training for Speech Recognition / [S. Schneider, A. Baevski, R. Collobert, M. Auli] // Interspeech. – 2019. – P. 1–9. DOI: 10.21437/interspeech.2019-1873
13. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale / [A. Babu, P. Wang, A. Mohamed, M. Karthik] // Interspeech. – 2021. – P. 2278–2282. DOI: 10.21437/Interspeech.2022-143
14. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units / [W. N. Hsu, B. Bolte, Y. Tsai et al.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2021. – Vol. 29. – P. 3451–3460.
15. Zbezhkhovska U. Exploring Challenges and Future Paths in Deepfake Audio Detection / U. Zbezhkhovska, O. Khapilin // Proceedings of the 5th Masters Symposium MS-AMLV-2024. – Lviv, Ukraine, March 29–30, 2024. – P. 1–10.
16. Spoofing Detection with DNN and One-class SVM for the ASVspoof 2015 Challenge / [J. A. V. López, M. P. Roddy, T. Kinnunen, Z. Tan] // Interspeech. – 2015. DOI:10.21437/Interspeech.2015-468
17. Amin T. B. Detecting Voice Disguise from Speech Variability: Analysis of Three Glottal and Vocal Tract Measures / T. B. Amin, J. S. German, P. Marziliano // Journal of the Acoustical Society of America. – 2013. – Vol. 20. DOI: 10.1121/1.4879257
18. Rodríguez-Ortega Y. A Machine Learning Model to Detect Fake Voice / Y. Rodríguez-Ortega, L. D. M. Ballesteros, D. Renza // International Conference on Applied Informatics. – 2020. Vol. 1277. – P. 3–13. DOI: 10.1007/978-3-030-61702-8_1
19. A Light CNN for Deep Face Representation with Noisy Labels / [X. Wu, R. He, Z. Sun, T. Tan] // IEEE Transactions on Information Forensics and Security. – 2018. – Vol. 13, № 11. – P. 2884–2896. DOI: 10.1109/TIFS.2018.2833032
20. Alzantot M. F. Deep Residual Neural Networks for Audio Spoofing Detection / M. F. Alzantot, Z. Wang, M. B. Srivastava // Interspeech. – 2019. – P. 1078–1082. DOI: 10.21437/Interspeech.2019-3174
21. Parasu P. Investigating Light-ResNet Architecture for Spoofing Detection under Mismatched Conditions / [P. Parasu, K. K. Park, Y. C. Lee, Y. Lee] // Interspeech. – 2020. – P. 1111–1115. DOI: 10.21437/Interspeech.2020-2039
22. Res2Net: A New Multi-scale Backbone Architecture / [S. H. Gao, M. Cheng, K. Zhao et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). – 2021. – Vol. 43, № 2. – P. 652–662. DOI: 10.1109/TPAMI.2019.2938758
23. End-to-end Anti-spoofing with RawNet2 / [H. Tak, J. Patino, M. Todisco, N. Evans] // ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2021. – P. 6369–6373. DOI: 10.1109/ICASSP39728.2021.9414234
24. Lai C. I. ASSERT: Antispoofing with Squeeze-Excitation and Residual Networks / C. I. Lai, S. H. Gao, Y. C. Chen // Interspeech. – 2019. – P. 1013–1017. DOI: 10.21437/Interspeech.2019-1794
25. End-to-end Spectro-temporal Graph Attention Networks for Speaker Verification Anti-spoofing and Speech Deepfake Detection / [H. Tak, J. Patino, M. Todisco, N. Evans] // Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. – 2021. – P. 1–8. DOI: 10.21437/ASVSPOOF.2021-1.
26. Detecting Spoofing Attacks using VGG and SincNet: Butomilia Submission to ASVspoof 2019 Challenge / [H. Zeinali, P. Mowlaee, J. Patino, N. Evans] // Interspeech. – 2019. – P. 1073–1077. DOI:10.21437/interspeech.2019-2892
27. Leveraging Positional-related Local-global Dependency for Synthetic Speech Detection / [X. Liu, J. Yang, H. Sun, L. Wang] // Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2023. – P. 1–5. DOI: 10.1109/ICASSP49357.2023.10096278
28. Deepfake Video Detection: YOLO-Face Convolution Recurrent Approach / [A. Ismail, M. S. Elpeltagy, M. S. Zaki, K. A. Eldahshan] // PeerJ Computer Science. – 2021. – Vol. 7. – P. 2–19. DOI:10.7717/peerj-cs.730
29. Hubálovský Š. Evaluation of Deepfake Detection using YOLO with Local Binary Pattern Histogram / [Š. Hubálovský, P. Trojovský, N. Bacanin, Kv. Venkatachalam] // PeerJ Computer Science. – 2022. – Vol. 8. DOI: 10.7717/peerj-cs.1086.
30. Sohan M. A Review on YOLOv8 and Its Advancements / M. Sohan, T. Sai Ram, C. V. Rami Reddy // Data Intelligence and Cognitive Informatics. – 2024. – P. 529–545. DOI: 10.1007/978-981-99-7962-2_39.
31. Redmon J. YOLOv3: An Incremental Improvement / J. Redmon, A. Farhadi // CoRR, arXiv preprint arXiv:1804.02767. – 2018. DOI: 10.48550/arXiv.1804.02767
32. Terven J. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS / J. Terven, D. M. Córdova-Esparza, J. A. Romero-González // Machine Learning and Knowledge Extraction. – 2023. – Vol. 5, № 4. – P. 1680–1716. DOI: 10.3390/make5040083
33. Real-Time Flying Object Detection with YOLOv8 / [D. Reis, J. Kupec, J. Hong, A. Daoudi] // arXiv preprint arXiv:2305.09972. – 2024. – P. 1–10.
34. Ruder S. An Overview of Gradient Descent Optimization Algorithms / S. Ruder // arXiv preprint arXiv:1609.04747. – 2017. – P. 1–14.
35. Ravanelli M. Speaker Recognition from Raw Waveform with SincNet / M. Ravanelli, Y. Bengio // 2018 IEEE Spoken Language Technology Workshop (SLT). – 2018. – P. 1021–1028. DOI: 10.1109/SLT.2018.8639585
36. Zbezhkhovska U. On Effectiveness and Generalization Capabilities of Deep Learning Models for Deepfake Audio Detection / U. Zbezhkhovska // Master Thesis. Ukrainian Catholic University, Faculty of Applied Sciences, Department of Computer Sciences. – Lviv, 2024. – 44 p.
37. t-EER: Parameter-Free Tandem Evaluation of Countermeasures and Biometric Comparators / [T. H. Kinnunen, K. A. Lee, H. Tak et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2023. – Vol. 46, № 5. – P. 2622–2637. DOI: 10.1109/TPAMI.2023.3313648
38. Müller N. M. Does Audio Deepfake Detection Generalize? / N. M. Müller, J. Partel, T. Kinnunen // Interspeech. – 2022. – P. 2783–2787. DOI: 10.21437/Interspeech.2022-108
39. Reimao R. FoR: A Dataset for Synthetic Speech Detection / R. Reimao, V. Tzerpos // 2019 IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD). – 2019. – P. 1–10. DOI: 10.1109/SpeD.2019.8878970
40. Frank J. WaveFake: A Data Set to Facilitate Audio DeepFake Detection / J. Frank, L. Schönherr // 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. – 2021. – P. 1–17. DOI: 10.5281/zenodo.5642694
41. Random Erasing Data Augmentation / [Z. Zhong, L. Zheng, G. Kang et al.] // The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20). – 2020. – P. 13001–13008.

DOI 10.15588/1607-3274-2025-1-14