12th INTERNATIONAL SCIENTIFIC CONFERENCE ON DEFENSIVE TECHNOLOGIES
OTEH 2026
Serbia, Tara, 21 - 24 October 2026

A COMPARATIVE EVALUATION OF SPECTROGRAM- AND BISPECTRUM- BASED TRANSFER LEARNING FOR DRONE AUDIO CLASSIFICATION

SARA MILANOVIĆ
Military Technical Institute, Belgrade, sara.milanovic@mod.gov.rs

MIODRAG VRAČAR
Military Technical Institute, Belgrade, mvracar566@gmail.com

IVAN POKRAJAC
Military Technical Institute, Belgrade, ivan.pokrajac@vs.rs

Abstract: *Acoustic signal analysis serves as an essential approach for detecting drones operated via optical fibers, which evade conventional radio-frequency detection methods. Nonetheless, achieving reliable acoustic-based detection poses notable challenges due to the pronounced sensitivity of classification models to environmental noise and low signal-to-noise ratios. While deep transfer learning has emerged as a powerful tool for audio classification, the choice of optimal time-frequency representation remains a critical factor in model performance. This paper presents a comparative evaluation of spectrogram- and bispectrum-based transfer learning for drone audio classification. Utilizing a comprehensive dataset of drone and non-drone acoustic signals, raw audio data is transformed into standard 2D spectrograms and higher-order bispectrum representations, which are then classified using three pre-trained models: ResNet-18, ResNet-50 and ViT Base 16. The results of this analysis suggest that higher-order spectral analysis provides a more robust and dependable feature representation in case of smaller model architectures, while larger networks struggle to generalize on the nonlinear features of the bispectrum and favor spectrogram representations.*

Keywords: *audio signal classification, drone classification, transfer learning, spectrogram, bispectrum.*

1. INTRODUCTION

The detection and classification of Unmanned Aerial Vehicles (UAVs), commonly known as drones, has become a matter of critical importance, namely for airspace surveillance and perimeter security. Existing counter-UAV systems employ several detection methods, including radar, radio-frequency (RF) monitoring, electro-optical and infrared imaging, and acoustic sensing. Each of these methods has its limitations – passive RF systems depend on the presence of detectable electromagnetic emissions and may be affected by interference from other wireless devices, whereas optical systems generally require an unobstructed view of the target and adequate illumination or atmospheric visibility [1]. These limitations motivate the investigation of complementary detection techniques.

The emergence of fibre-optic-controlled UAVs further illustrates the limitations of systems that depend on detecting or disrupting wireless command links. In such platforms, command and video signals are transmitted through a physical fibre-optic cable rather than a conventional radio link. Consequently, RF detectors designed to identify command-and-control transmissions may be ineffective [2]. Acoustic sensing is therefore of interest as a passive and comparatively low-cost method that does not depend on the UAV's communication protocol. It instead exploits the characteristic sounds generated by rotating propellers, electric motors, and associated mechanical components [3]. Its practical performance, however, is affected by propagation distance, microphone characteristics, wind, reverberation, and competing environmental sounds [4].

Machine Learning (ML)-based acoustic UAV detection commonly begins by converting a time-domain recording into a representation that exposes its spectral and temporal structure. Frequently used representations include short-time Fourier transform spectrograms, Mel-frequency spectrograms, filterbank energies, and Mel-frequency cepstral coefficients (MFCCs) [1], [3], [5]. A magnitude or power spectrogram describes the local distribution of signal energy across time and frequency, but it omits the phase of the complex STFT coefficients. In addition, additive noise contributes energy to the same time-frequency representation and may obscure the tonal and harmonic components associated with rotor and motor operation.

Higher-order spectral analysis offers a complementary representation of these signals. In particular, the bispectrum describes third-order statistical dependencies and can reveal quadratic phase coupling among frequency components [6], [7]. Such coupling may provide information about harmonic and nonlinear interactions that is not represented explicitly in an ordinary power spectrum. Moreover, for additive Gaussian noise that is statistically independent of the signal, third-order cumulants vanish in

Corresponding author: S. Milanović;
E-mail: sara.milanovic@mod.gov.rs
DOI: 10.5937/OTEH26059M

expectation. Bispectral estimation can therefore reduce the expected contribution of Gaussian noise [6], [7]. Most frequently used methods for estimating this bispectral structure include parametric approaches based on autoregressive moving-average (ARMA) models [8].

The selection of a suitable classifier is also influenced by the quantity and diversity of the available training data. Convolutional Neural Networks (CNNs), including residual architectures such as ResNet, incorporate local spatial structure through convolution and have been widely adopted for time-frequency image inputs [9]. Transformer architectures, including the Vision Transformer and the Audio Spectrogram Transformer, use self-attention to model relationships between patches or time-frequency regions over longer ranges [5], [10]. However, the original success of vision transformers depended substantially on large-scale pretraining, and training such models from scratch can be difficult when only a limited dataset is available. Data-efficient training, transfer learning, augmentation, and regularization can mitigate this limitation [10], [11]. Thus, model performance in small datasets depends not only on architecture capacity, but also on initialization and training strategy.

2. LITERATURE

Acoustic UAV detection has predominantly been formulated as an environmental sound-classification problem using short-time spectral representations. Jeon et al. [12] compared Gaussian mixture models, CNNs, and Recurrent Neural Networks (RNNs) using MFCC and Mel-spectrogram features, including evaluation with recordings made in an urban street environment. Bernardini et al. [13] adopted a more traditional pipeline, combining temporal and spectral descriptors—such as short-time energy, spectral centroid, spectral roll-off, and MFCCs—with a support vector machine (SVM) model. Al-Emadi et al. [14] evaluated CNN, RNN, and Convolutional Recurrent Neural Network (CRNN) architectures using spectrogram representations and investigated synthetic data generation as a means of enlarging the available UAV-audio dataset. Other studies have explored model and feature fusion: Casabianca and Zhang [15] combined the outputs of multiple deep networks operating on Mel-spectrograms, whereas Dong et al. [16] fused decisions obtained from Log-Mel-spectrogram and MFCC-based classifiers. Collectively, these studies demonstrate the feasibility of learning drone acoustic signatures from time-frequency representations.

Nevertheless, the robustness of acoustic UAV detectors remains sensitive to propagation distance, environmental noise, microphone characteristics, UAV type, and differences between training and deployment conditions. Tejera-Berengué et al. [4] explicitly examined the effects of distance and recording environment on several conventional and deep-learning classifiers. More recent work has similarly emphasized cross-dataset generalization and the value of pretraining and noise augmentation [17]. Thus, it is safe to say that many published results are obtained from relatively limited, augmented, or closely matched datasets, and that reliable performance in previously unseen acoustic environments remains an open challenge.

As mentioned, most existing acoustic UAV classifiers rely on second-order signal representations (power spectra, STFT spectrograms, Mel-spectrograms), while bispectral representations have been investigated primarily for radar and RF signal classification rather than acoustic detection. For example, Liu et al. [18] combined bispectral slices with a neural classifier to recognize low signal-to-noise (SNR) UAV radar-signal modulations. In the available literature analyzed for this study, no peer-reviewed work was found that combines UAV acoustic bispectra with pretrained vision architectures. Parametric bispectrum estimation is itself well established. Raghuveer and Nikias developed an autoregressive approach in which model parameters are estimated from third-order recursion equations, providing a parametric alternative to conventional Fourier-based bispectrum estimators [5].

3. METHODOLOGY

In this section, we describe the framework utilized to classify drone audio signals and evaluate the performance of the model. Section 3.1 describes the dataset used in this study, along with the necessary preprocessing steps. Following the data description, Section 3.2 introduces the feature extraction pipeline. We give a brief explanation of spectrograms and bispectral representations of audio signals, used to capture temporal, spectral and higher-order statistical variations. In Section 3.3, the process of model training and evaluation is outlined, focusing on the method of transfer learning and the implemented cross-validation evaluation strategy.

3.1 Dataset

The dataset used for this research was taken from the HuggingFace hub [19]. It consists of ~180k .wav files with sampling rate of 16kHz, and each file belongs to one of the two categories: DRONE or NO_DRONE. Because of the substantial volume of the dataset and the resulting computational costs, we extracted an approximately balanced subset for the purpose of this study.

Since the original files are of varying lengths, they were segmented into 1-second snippets to satisfy the fixed-input requirements of neural network training process. This can result in the appearance of snippets containing no valuable information, which is why the resulting dataset was filtered using a percentile-based root mean square (RMS) energy threshold. With this, we acquired a subset consisting of ~3500 audio signals.

As it can be concluded from the class names, the DRONE class includes audio signals which contain various drone sounds in different operating environments. The samples were aggregated from multiple data sources, primarily capturing the acoustic profiles of various multirotor UAVs, with commercial quadcopters being the most dominant. On the other hand, the audio files in the NO_DRONE class represent a variety of sounds, including traffic, animal noises, chatter of people, etc. and were specifically pulled from environmental audio datasets. In some of the files, multiple different sounds are combined. For example, audio files from DRONE class may include distant traffic,

wind and outdoor ambient noise in the background.

3.2 Feature extraction

One of the most important steps before starting the ML model training process is selecting the right features to appropriately represent the data we want to classify. When working with audio files, there is a possibility of extracting features from the time domain (mean, standard deviation, zero-crossing rate, etc.), but this has ultimately proven to be inferior to the frequency-domain approach.

Figures 1 and 2 show examples of spectra of signals from our two different classes. As we can see in the DRONE signal spectrum, there is a series of regularly spaced, distinct peaks which correspond to the harmonic comb structure. This pattern can be attributed to the drone's electric motors and rotating propeller blades.

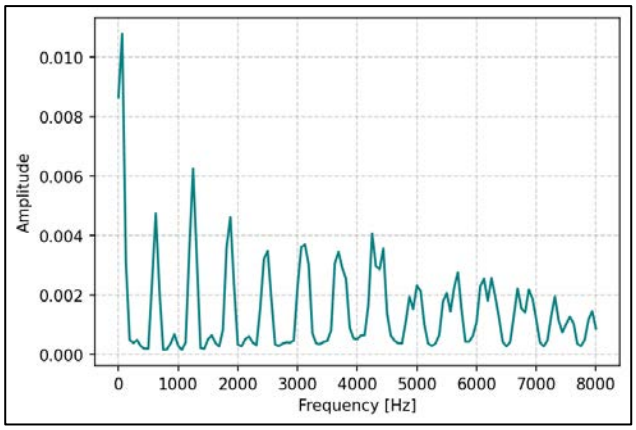


Figure 1. Spectrum of a DRONE signal

The difference between these two types of signals is clear when comparing figures 1 and 2. The NO_DRONE signal's energy is concentrated in the lower frequencies and undergoes a rapid decay past 1000Hz. The lack of mechanical periodicity here is obvious and the spectrum's appearance is highly characteristic of ambient background noise.

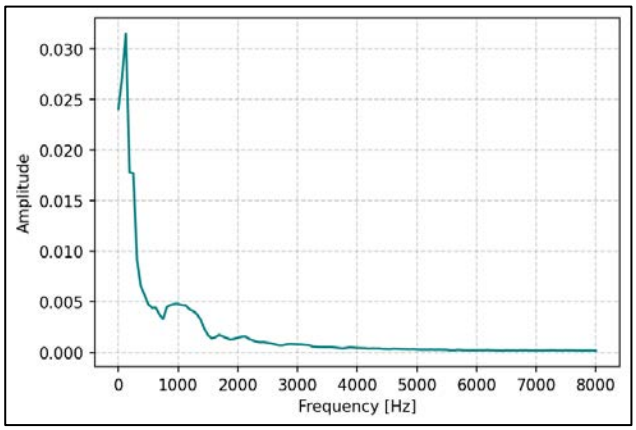


Figure 2. Spectrum of a NO_DRONE signal

Spectrogram is one of the most commonly used time-frequency representations of a signal. We use the Discrete Short-Time Fourier Transform (DSTFT) to calculate the frequency-domain representation of an audio file, given by the formula:

$$X[m,k] = \sum_{n=0}^{N-1} x[n+mR]w[n]e^{-j\frac{2\pi}{N}kn} \quad (1)$$

In equation (1), $X[m,k]$ is the complex-valued STFT coefficient at time index m and frequency bin k, $x[n+mR]$ is the $(n+mR)$-th sample of the original audio sequence, $w[n]$ is the window function, R is the hop size, and N is the total number of samples. The spectrogram can then be calculated as:

$$S[m,k] = |X[m,k]|^2 \quad (2)$$

While they are highly effective for extracting stationary characteristics, they cannot fully capture complex, non-linear dynamics often present in audio signals.

Figures 3 and 4 show examples of spectrograms of signals from the two different classes.

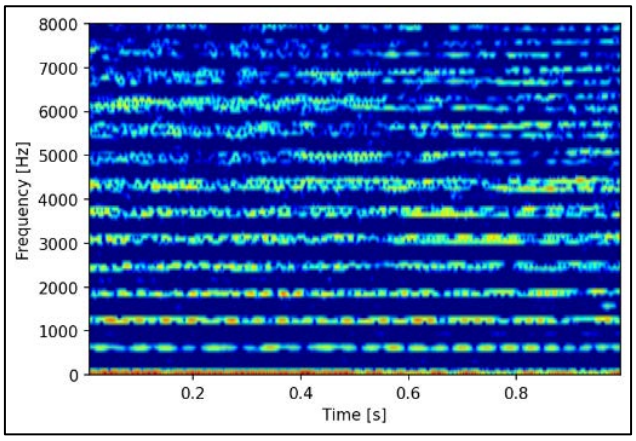


Figure 3. Spectrogram of a DRONE signal

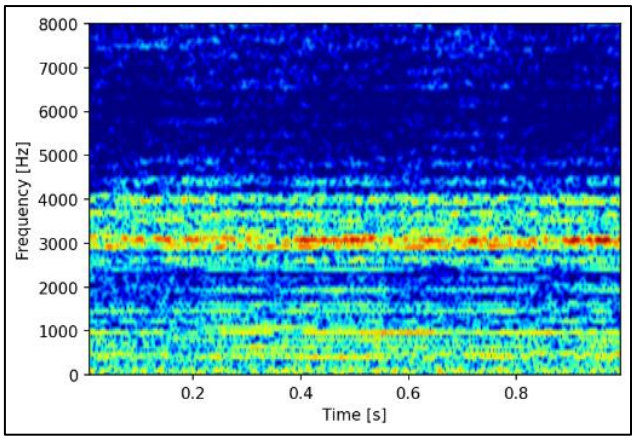


Figure 4. Spectrogram of a NO_DRONE signal

From these two representations, we can draw the same conclusions as we did when looking at figures 1 and 2. In the case of DRONE signal, we can see multiple "stripes" in the spectrogram which are the result of drone's motors and rotating blades. On the other hand, the NO_DRONE signal spectrogram has no such patterns and its values are noticeably lower for frequencies past ~4kHz.

Because of spectrogram's limitations, we calculated the bispectrum (third-order cumulant spectrum). The advantage of the bispectrum is in its ability to preserve phase information and detect non-linear interactions

between frequency components. Mathematically, the bispectrum is defined as the Fourier transform of its third-order cumulant sequence. Again, we use the DFT to practically calculate the bispectrum of a signal:

$$B[k_1, k_2] = E\left[X[k_1]X[k_2]X^*[k_1 + k_2]\right] \qquad (3)$$

where $X[k1]$ and $X[k2]$ are the DFT coefficients at the frequency bins $k1$ and $k2$, respectively, X^* is the complex conjugate of the DFT coefficient at the sum $[k_1 + k_2]$ frequency bin, and E is the expectation operator. Because the resulting bispectrum is complex-valued, we calculate the absolute magnitude to extract a real-valued matrix suitable for neural network training.

Alternatively, the bispectrum can be calculated using Autoregressive Moving Average (ARMA) coefficients, and this is a parametric method that we used for this research. The method relies on modeling the audio signal as an ARMA process, estimating the coefficients and calculating the transfer function. Without going into the mathematics of this approach, the bispectrum is calculated as:

$$B[k_1, k_2] = \gamma_3 H(k_1)H(k_2)H^*(-k_1 - k_2) \qquad (4)$$

In the previous equation, H and γ_3 are the transfer function and parameter obtained from the ARMA estimation. Determining the ARMA model order is tricky, but if done right, the resulting bispectrum will have higher frequency resolution and less smoothing issues (which are inherent to windowing functions) than is the case when calculating spectrograms.

For the research done in this paper, we used MatLAB to perform automatic ARMA model order selection [20] and the corresponding bispectrum matrices calculation, which we then loaded into Python prior to the model training. Spectrograms were calculated directly in Python. Figures 5 and 6 show examples of the resulting bispectra.

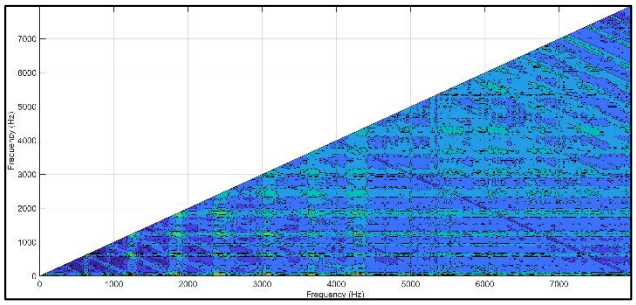


Figure 5. Bispectrum of a DRONE signal

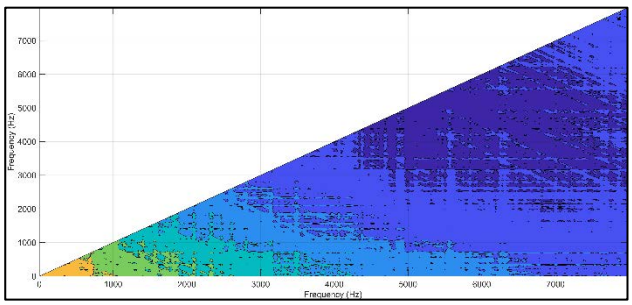


Figure 6. Bispectrum of a NO_DRONE signal

As can be seen in figures 5 and 6, because of bisepctrum's symmetrical nature and in order to lower the computational costs, we only kept the right-triangular matrix information for each bispectrum. Also, the bispectral magnitudes were converted to a logarithmic scale, which provided better results than linear scaling.

In these figures we can notice the same acoustic signatures we observed by analyzing the spectra. When looking at the bispectrum of a DRONE signal, we notice the intersecting vertical and horizontal lines which effectively form a grid pattern. This is the result of drone's deterministic, mechanical parts – rotating motors and blades.

The NO_DRONE signal's bispectrum shows strong phase coupling only in the low-frequency range. Signals from this class include various environmental sounds that are highly localized in both time and frequency, which explains why they are present as small, isolated phase coupling clusters in the bispectrum. As they lack the deterministic nature of drone signals, it makes perfect sense that there are no grid-like patterns visible, with only the dark blue color dominating the high-frequency range.

3.3 Model training and evaluation

As mentioned before, we opted for transfer learning with three pretrained models for our analysis: ResNet-18, ResNet-50 [9] and ViT Base 16 [10]. Instead of training a completely new neural network from scratch, transfer learning refers to using an already trained network and slightly adjusting it for a specific task. With this approach, we can be sure that the model has already learned common important features and we only need to fine-tune it, which also significantly reduces the amount of parameters it needs to learn and the resources it uses during that process.

Expectedly, this then poses a question: how many layers should we retrain (fine-tune) when using transfer learning? The models we selected vary significantly in the number of trainable parameters. ResNet-18 and ResNet-50 consist of roughly 12M and 26M parameters, respectively. On the other hand, ViT Base 16 is considerably larger, including around 85M trainable parameters. Considering the size of our dataset, we opted for the next approach: for ResNet-18 and ResNet-50, we made only deeper residual blocks trainable (layer3 and layer4 in PyTorch implementation). In case of ViT Base 16 model, the final six encoder blocks (blocks 6-11) were unfrozen for fine-tuning. The final fully connected layer is the same in all three cases — it includes a layer mapping the backbone features to a 128-dimensional space, followed by a ReLU activation function and a dropout layer ($p = 0.4$) to combat overfitting. For the optimizer we used AdamW, with learning rates ranging from 10^{-3} to 10^{-5}, depending on the depth of the layers.

One of the biggest concerns when it comes to training neural networks is their resulting generalization ability. It is not correct to focus solely on training accuracy and loss, because of possible overfitting. That is why we split the entire dataset into training and validation subsets and monitor the model's performance on both. To get an even better idea of our model's ability to generalize, we can use k-fold cross-validation. This method involves splitting the dataset into k subsets, using one subset for validation and the remaining $(k - 1)$ subsets for training. This process is

repeated k times and the resulting metrics are averaged. In this paper, we use 10-fold cross-validation (90% of the data is used for training and 10% for validation) to calculate our model's accuracy on the OOF (out-of-fold) samples (samples that were in the validation subset in each fold). Furthermore, the technique called Early Stopping is used during training in each fold. We observe validation loss with the patience parameter set to 3, meaning that if the loss did not achieve a new minimum after 3 epochs, the training process is terminated and the last best model weights are recorded and used for evaluation.

4. RESULTS

In this section, we evaluate the classification performances of our three selected models across two distinct feature extraction paradigms: the non-parametric STFT spectrogram and the parametric ARMA bispectrum. All models were trained and evaluated using the partial fine-tuning technique and 10-fold cross-validation, as described in the previous section. Because we used cross-validation for model training and evaluation, classification performance is quantified with mean accuracy (along with standard deviation), calculated on the OOF (out-of-fold) samples. Table 1 presents this metric for all mentioned experimental configurations.

Table 1. Classification results

	ResNet-18	ResNet-50	ViT Base16
Spectrogram	96.66 (±1.20)	96.46 (±0.89)	95.50 (±0.92)
Bispectrum	**96.92 (±0.78)**	96.05 (±1.19)	93.20 (±1.93)

We can draw several interesting conclusions looking at the results presented in Table 1. Firstly, ResNet-18 achieved the best classification performance, with bispectrum representations as inputs. The lowest standard deviation also makes this model the most consistent. This is particularly interesting because ResNet-18 has the smallest architecture here.

On the other hand, while it yields worse results than ResNet-18, ResNet-50 outperforms ViT Base 16 for both feature extraction methods. Especially intriguing is the fact that both ResNet-50 and ViT Base 16 achieve higher accuracy with spectrogram inputs. It could be argued that one of the reasons for these outcomes is the size of the dataset. An insufficient number of samples can cause models with large number of parameters (like ResNet-50 and especially ViT) to start learning specific details in training data that are not descriptive and/or represent noise. This ultimately leads to worse generalization ability and can only be remedied by increasing the dataset size. The explanation why larger models favor spectrograms in this experiment is not easy to explain. Spectrograms preserve the temporal information of the signal, while the ARMA bispectrum completely eliminates the time component in favor of focusing on frequency interactions. Because of their complex architectures and massive parameter numbers, large models may end up struggling trying to generalize on the bispectrum's nonlinear features.

We can take a look at the confusion matrix and ROC curve for the best-performing model, ResNet-18. Figure 7 shows the confusion matrix for said model.

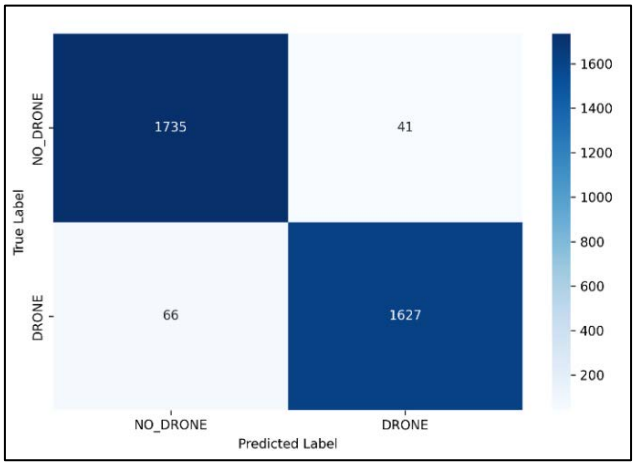


Figure 7. Classification matrix for ResNet-18

The matrix was obtained by aggregating all the OOF predictions, which is why the number of samples in the matrix is equal to the dataset size. As can be seen from the figure, classification performance is roughly symmetrical; the model achieves high precision and recall, indicating that there is no bias toward one specific class. The number of False Negative cases is higher than the number of False Positives, which is unfavorable for this specific task of drone classification.

Figure 8 shows the ROC curve for the best-performing model. The model achieved an Area Under the Curve (AUC) of 0.994. Because this metric was calculated based on out-of-fold (OOF) predictions across all ten folds, it provides a robust estimate of the model's classification performance and generalization ability.

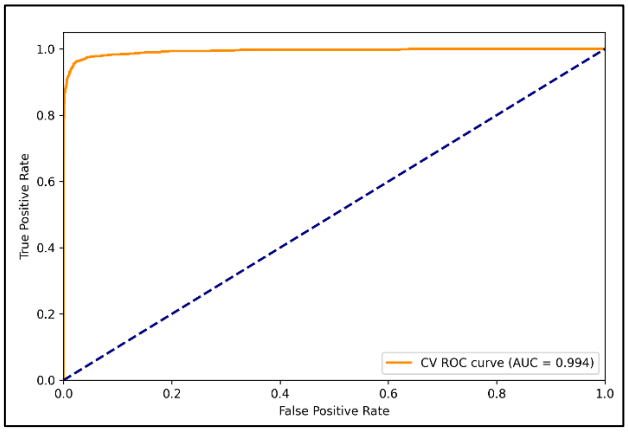


Figure 8. ROC curve for ResNet-18

5. CONCLUSION

In this paper, we performed a comparative evaluation of various transfer learning paradigms, focusing on audio drone classification. We utilized three different model architectures – ResNet-18, ResNet-50 and ViT Base 16. The networks were cross-examined across two distinct feature spaces: non-parametric STFT spectrograms and parametric ARMA bispectra.

The experiments revealed some unexpected results. The smallest model, ResNet-18, consistently outperformed

larger models, in both feature spaces. This best-performing model yields a ROC curve with the AUC of 0.994, with symmetrical classification performance. Non-parametric STFT spectrograms proved more valuable for larger networks, which is especially evident in the case of ViT Base 16. This model achieves a significantly higher accuracy when accepting spectrograms as inputs than is the case with bispectrum data.

It should be mentioned that the primary limitation of current research is the size of the dataset used. For large model architectures (notably ViT Base 16 in this case), data count is of critical importance. Consequently, the logical next step in this research would be the expansion of the dataset. With this, new experiments can be performed, which would help in determining whether large deep learning models can learn to better generalize on ARMA-calculated bispectrum inputs.

Based on the currently obtained results, another research direction involves investigating the usage of smaller model architectures (because of their superior performance) and performing testing with audio signals of drones operating in real-world environments.

References

[1] U. Seidaliyeva, L. Ilipbayeva, K. Taissariyeva, N. Smailov, and E. T. Matson, "Advances and Challenges in Drone Detection and Classification Techniques: A State-of-the-Art Review," *Sensors*, vol. 24, no. 1, art. 125, 2024.

[2] A. Braszko, *Fiber Optic Drones: Posing a Significant C-UAS Challenge*, Center for Army Lessons Learned, No. 25-1046, 2025.

[3] S. Al-Emadi, A. Al-Ali, A. Mohammad, and A. Al-Ali, "Audio Based Drone Detection and Identification Using Deep Learning," in *2019 15th International Wireless Communications and Mobile Computing Conference*, pp. 459–464, 2019.

[4] D. Tejera-Berengué, F. Zhu-Zhou, M. Utrilla-Manso, R. Gil-Pita, and M. Rosa-Zurera, "Analysis of Distance and Environmental Impact on UAV Acoustic Detection," *Electronics*, vol. 13, no. 3, art. 643, 2024.

[5] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proceedings of Interspeech 2021*, pp. 571–575, 2021, doi: 10.21437/Interspeech.2021-698.

[6] C. L. Nikias and M. R. Raghuveer, "Bispectrum Estimation: A Digital Signal Processing Framework," *Proceedings of the IEEE*, vol. 75, no. 7, pp. 869–891, 1987.

[7] J. M. Mendel, "Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoretical Results and Some Applications," *Proceedings of the IEEE*, vol. 79, no. 3, pp. 278–305, 1991.

[8] M. R. Raghuveer and C. L. Nikias, "Bispectrum Estimation: A Parametric Approach," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 5, pp. 1213–1230, 1985.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[10] A. Dosovitskiy et al., "An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021.

[11] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training Data-Efficient Image Transformers and Distillation through Attention," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 10347–10357, 2021.

[12] S. Jeon et al., "Empirical Study of Drone Sound Detection in Real-Life Environment with Deep Neural Networks," in *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)*, 2017.

[13] A. Bernardini, F. Mangiatordi, E. Pallotti, and L. Capodiferro, "Drone Detection by Acoustic Signature Identification," *Electronic Imaging*, pp. 60–64, 2017.

[14] S. Al-Emadi, A. Al-Ali, and A. Al-Ali, "Audio-Based Drone Detection and Identification Using Deep Learning Techniques with Dataset Enhancement through Generative Adversarial Networks," *Sensors*, vol. 21, no. 15, art. 4953, 2021.

[15] P. Casabianca and Y. Zhang, "Acoustic-Based UAV Detection Using Late Fusion of Deep Neural Networks," *Drones*, vol. 5, no. 3, art. 54, 2021.

[16] Q. Dong, Y. Liu, and X. Liu, "Drone Sound Detection System Based on Feature Result-Level Fusion Using Deep Learning," *Multimedia Tools and Applications*, vol. 82, pp. 149–171, 2023.

[17] J. Reuter, H. Ohlenbusch, and S. Rollwage, "Improving Acoustic Drone Detection Generalization through Pretraining and Data Augmentation," arXiv:2605.31329, 2026.

[18] Liu et al., "Modulation Recognition of Low-SNR UAV Radar Signals Based on Bispectral Slices and GA-BP Neural Network," *Drones*, vol. 7, no. 7, art. 472, 2023.

[19] GeronimoBasso. (n.d.).: *Drone audio detection samples (DADS) [Dataset]*. HuggingFace. https://huggingface.co/datasets/geronimobasso/drone-audio-detection-samples.

[20] P. M. T. Broersen. ARMAsel for Irregular or Missing Data (https://www.mathworks.com/matlabcentral/fileexchange/18429-armasel-for-irregular-or-missing-data), MATLAB Central File Exchange. Retrieved August 5, 2026.