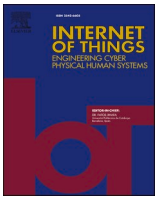
Internet of Things 23 (2023) 100879

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.sciencedirect.com/journal/internet-of-things



# Radio frequency fingerprint-based drone identification and classification using Mel spectrograms and pre-trained YAMNet neural

Kamel K. Mohammed [a,1], Eman I.Abd El-Latif [b,1], Noha Emad El-Sayad [c,1], Ashraf Darwish [d,1], Aboul Ella Hassanien [e,*,1]

[a] Centre for Virus Research and Studies, Al-Azhar University, Cairo, Egypt
[b] Faculty of Science, Benha University, Benha, Egypt
[c] Electronics and Communications Department, Faculty of Engineering, Horus University-Egypt (HUE), Egypt
[d] Faculty of Science, Helwan University, Cairo, Egypt
[e] Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt

ARTICLE INFO

*Keywords:*
Drone identification
Classification
IoT
Radio-frequency fingerprints
Mel spectrograms
Neural networks
YAMNet
Transfer learning
Audio signals

ABSTRACT

The convergence of drones with the Internet of Things (IoT) has paved the way for the Internet of Drones (IoD), an interconnected network of drones, ground control systems, and cloud infrastructure that enables enhanced connectivity, data exchange, and autonomous operations. The integration of drones with the IoD has opened up new possibilities for efficient and intelligent aerial operations, facilitating advancements in sectors such as logistics, agriculture, surveillance, and emergency response This paper introduces a novel approach for drone identification and classification by extracting radio frequency (RF) fingerprints and utilizing Mel spectrograms as distinctive patterns. The proposed approach converts RF signals to audio signals and leverages Mel spectrograms as essential features to train neural networks. The YAMNet neural network is employed, utilizing transfer learning techniques, to train the dataset and classify multiple drone models. The initial classification layer achieves an impressive accuracy of 99.6% in distinguishing between drones and non-drones. In the subsequent layer, the model achieves 96.9% accuracy in identifying drone types from three classes, including AR Drone, Bebop Drone, and Phantom Drone. At the third classification layer, the accuracy ranges between 96% and 97% for identifying the specific mode of each drone type. This research showcases the efficacy of Mel spectrogram-based RF fingerprints and demonstrates the potential for accurate drone identification and classification using pre-trained YAMNet neural networks.

## 1. Introduction

Unmanned aerial vehicles (UAVs), commonly referred to as drones, have become increasingly prevalent in various sectors, offering substantial benefits in terms of efficiency, convenience, and cost-effectiveness [1–3]. However, as their usage expands, concerns regarding their potential misuse and security implications have also emerged [4,5]. The integration of our drone identification and

* Corresponding author.
*E-mail address:* aboitcairo@cu.edu.eg (A.E. Hassanien).
[1] Scientific Research Group in Egypt (SRGE), www.egyptscience.net

https://doi.org/10.1016/j.iot.2023.100879

Available online 19 July 2023
2542-6605/© 2023 Elsevier B.V. All rights reserved.

classification model with the IoD ecosystem extends the potential applications and benefits. The identified drones can be seamlessly connected to the broader network, enabling real-time monitoring [6], coordinated operations [7], and data-driven decision-making [8]. Furthermore, the extracted RF fingerprints can contribute to the development of comprehensive drone registries and airspace management systems within the IoD, ensuring safe and efficient drone operations.

Drone identification and classification based on RF fingerprints involves capturing and analyzing the unique radio frequency signals emitted by drones during their operation [9–11]. Traditional methods for drone identification have primarily relied on visual cues, such as image or video analysis [12–14]. However, these approaches can be limited in scenarios where drones are visually obstructed or when distinguishing between similar drone models is challenging. RF-based identification techniques provide an alternative means to overcome these limitations and offer reliable identification capabilities.

To address these concerns, reliable and accurate drone identification and classification techniques are essential. In this paper, we present a novel approach for drone identification and classification by extracting radio frequency (RF) fingerprints, and to achieve accurate identification and classification, we leverage the YAMNet neural network, which has demonstrated exceptional performance in audio analysis tasks. With a transfer learning approach, we adapt the pre-trained YAMNet model to our specific drone identification and classification task. By leveraging the model's knowledge acquired from a large-scale audio dataset, we expedite the training process and enhance the accuracy of our system.

The proposed model begins by converting RF signals into audio signals, enabling the utilization of audio analysis techniques for feature extraction. Specifically, we employ Mel-spectrograms, which are widely used for audio signal analysis, to capture the frequency and intensity patterns unique to each drone. The extracted Mel-spectrograms serve as distinctive fingerprints that capture the inherent characteristics of drone RF signals. The main contributions of this paper can be summarized as follows:

1 Mel Spectrogram-based RF Fingerprinting: The paper introduces a novel approach that utilizes Mel spectrograms as unique RF fingerprints for drone identification. This method effectively captures the frequency energy distribution of drone signals over time, providing important features for distinguishing and classifying drones.
2 Pre-trained YAMNet Neural Network: The paper adapts a pre-trained YAMNet neural network for the classification of drones based on the extracted RF fingerprints. Leveraging transfer learning techniques, the YAMNet network demonstrates high performance in accurately identifying different drone types and modes. The proposed model achieves impressive classification accuracy. In the first classification layer, the model achieves 99.6% accuracy in distinguishing between drones and non-drones. In the subsequent layers, it achieves 96.9% accuracy for identifying drone types and 96–97% accuracy for determining the mode of each type.
3 Effective Representation of Human Perception: The use of the Mel frequency scale in the spectrogram generation captures the non-linear characteristics of human auditory perception. This provides a more reliable representation of the frequencies typically perceived by humans and enhances the model's ability to differentiate between drone signals.

These contributions collectively advance the field of drone identification and classification, offering a robust and accurate method for differentiating between drones and determining their specific types and modes. The proposed approach holds significant potential for enhancing drone surveillance and security systems in various domains.

The remainder of this paper is organized as follows. Section 2 introduces the previous research on the topic. Section 3 presents a detailed analysis and description of the data used in this study. The proposed algorithm and YAMNet model are then introduced in Section 4. The proposed Mel Spectrogram-based fingerprinting describes in Section 5. In Section 6, the experimental results are evaluated. Finally, the study is concluded in Section 7.

## 2. Related work

Drone identification and detection have garnered significant attention in recent years due to the increasing presence of drones in various domains. This section presents an overview of the existing related work on drone identification through different techniques such as video analytics, radar cross-section, and RF signal.

In the domain of drone identification, the existing literature on drone identification notable study by Aker et al. [15] introduced the concept of using image processing and video analytics to train machine learning algorithms for classifying drones. These algorithms analyze video sequences captured by various types of cameras to extract features for various classifications, such as drone category, differentiating between drones and birds, and evaluating the presence of payloads that affect the target's radar cross section (RCS) [16].

However, the fundamental constraint of this system is the small size of drone targets, which can easily blend into the background or resemble birds. Furthermore, environmental variables, such as low ambient light, fluctuating lighting, shadows, and lens occlusion, pose obstacles for camera-based systems, leaving the system inoperable. Infrared cameras are necessary in dark places or at night, however, they frequently have lesser resolution and range and are more expensive. Given these constraints, video-based techniques are best successful in good weather and at short distances. They are especially effective for validating and categorizing drones after first detection.

Taha et al. [17] presented a comprehensive review of video-based techniques for drone detection. However, they noted that there is relatively less survey of the literature on radar techniques for drone detection. Before discussing the details of radar approaches, specifically those based on active sensors (radar), it is worth mentioning that acoustic sensors have also been explored as a potential technology for identifying the presence of drones using statistical array processing techniques.

Acoustic sensors utilize sound presented by Anwar et al. [18,19] waves to detect and locate objects, including drones. By employing

an array of microphones, it becomes possible to analyze the received audio signals and extract relevant information. Statistical array processing techniques, such as beam forming and direction of arrival estimation, are commonly used to process the audio data and identify the presence of drones. One advantage of acoustic sensors is that they can operate in different environmental conditions, including low-light or visually obscured scenarios. Additionally, they can provide valuable supplementary information when used in conjunction with video or radar-based detection systems.

On the other hand, acoustic sensors have certain limitations. They are susceptible to ambient noise and may struggle to distinguish between drone sounds and other background sounds, which can lead to false alarms or missed detections. Moreover, their range and accuracy may be influenced by factors like wind, temperature, and atmospheric conditions. While video-based techniques have received more attention in the literature, radar-based approaches offer their own set of advantages and challenges. Radar systems use radio waves to detect objects, including drones, by analyzing the reflected signals. These systems can provide information about the drone's position, velocity, and size.

Radar-based drone detection has advantages such as long-range capabilities, resistance to visual obstructions, and the ability to operate in various weather conditions. They can also offer continuous monitoring and real-time tracking of multiple targets. However, radar systems can be expensive, and their installation and maintenance require specialized knowledge.

In summary, Taha et al. [17] highlighted the prevalence of video-based techniques in the literature on drone detection. However, they acknowledged the need for more exploration and research on radar techniques. Acoustic sensors have also been considered as a viable alternative, utilizing statistical array processing techniques to detect drones based on sound. Each technology, whether video-based, radar-based, or acoustic, has its own strengths and limitations, and further research is necessary to improve and integrate

**Table 1**
Comparison of Pros and Cons of Different UAV Classification Techniques.

| Technique | Pros | Cons |
|---|---|---|
| IMAGE PROCESSING AND VIDEO ANALYTIC [15, 16] | 1 Accurate classification of drones based on visual data.<br>2 Automation of the classification process, reducing manual intervention.<br>3 Scalability for processing and analyzing large volumes of image and video data.<br>4 Adaptability to different drone models, sizes, and flight patterns. | 1 Limited generalization if the training data is not diverse of real-world scenarios.<br>2 Vulnerability to variations in lighting, or weather conditions.<br>3 Requirement of significant computational resources for training and running the algorithms.<br>4 Privacy concerns related to the analysis of visual data captured in public or private spaces. |
| RADAR [17] | 1 Effective detection and classification of drones based on their radar cross-section signatures.<br>2 Reduced dependence on visual conditions, such as lighting or weather.<br>3 Robustness in detecting drones even in environments with visual obstructions like buildings.<br>4 Potential for long-range detection and classification of drones, providing an extended surveillance capability. | 1 Limited classification accuracy if the radar cross-section signatures of different drones overlap or if drones exhibit low radar cross-sections.<br>2 Higher susceptibility to false positives or false negatives due to the presence of clutter or interference in the radar environment.<br>3 Higher complexity and cost.<br>4 Limited ability to gather detailed visual information about the drone. |
| AUDIO SIGNALS [18,19] | 1 Effective detection and classification of drones based on their unique audio signatures or acoustic features.<br>2 Potential for long-range detection and classification of drones, as sound can propagate over long distances.<br>3 Lower complexity and cost compared to radar or visual-based methods. | 1 Limited classification accuracy if the audio signatures of different drone models overlap or if drones exhibit low sound levels.<br>2 Susceptibility to false positives or false negatives due to environmental noise or similar acoustic patterns from non-drone sources.<br>3 Difficulty in gathering detailed visual information about the drone (e.g., shape, size, and payload) solely from audio signals. |
| RF FINGERPRINTS [20] | 1 Effective detection and classification of drones based on their unique RF fingerprints or electromagnetic emissions.<br>2 Independent of visual or acoustic conditions, allowing for detection in various environments regardless of lighting or noise levels.<br>3 Potential for long-range detection and classification of drones, as RF signals can propagate over considerable distances.<br>4 Lower complexity and cost compared to some other methods. | 1 Limited classification accuracy if the RF fingerprints of different drone models overlap or if drones exhibit low RF emissions.<br>2 Susceptibility to false positives or false negatives due to environmental RF noise or similar RF patterns from non-drone sources.<br>3 Reliance on clear RF signals, making the method less effective in congested RF environments or areas with significant RF interference. |
| MEL-SPECTROGRAMS [21–24] | 1 Effective representation of audio data by converting it into a visual form that captures frequency and time information.<br>2 Mel-spectrograms provide a comprehensive and detailed representation of audio signals, enabling accurate classification of drones based on their acoustic features.<br>3 Independence from visual conditions, allowing for detection in low light or obscured environments.<br>4 Compatibility with various machine learning algorithms, making it easier to train and classify drones using a wide range of models. | Limited classification accuracy if the Mel-spectrograms of different drone models overlap or if drones exhibit similar acoustic characteristics.<br>Susceptibility to false positives or false negatives due to background noise or interference that can affect the spectrogram representation.<br>Reliance on clear audio signals, making the method less effective in noisy or congested acoustic environments.<br>Difficulty in gathering detailed visual information about the drone solely from Mel-spectrograms. |

these approaches for effective drone detection. Radar-based RF signals provide weather-resistant and long-range coverage, making them suitable for primary drone detection, with radar sensors being the most important active RF-based device, while passive technologies are also relevant.

In the domain of drone identification, researchers have recognized the potential of RF fingerprints as a key feature for accurate classification. A notable study by Wei Nie et al. in [20] introduced the concept of applying RF Fingerprints to detect and identify drones. The technique initially conducts drone detection and after that, three different equations are extracted as a drone fingerprint. However, their work mainly focused on basic drone detection and did not explore the potential of more advanced techniques such as spectrogram analysis.

To further enhance the identification accuracy, several researchers have incorporated spectrogram analysis into the pipeline [21, 22,23]. Ezuma et al. in [24] proposed the use of Mel-spectrograms, a representation of audio signals that highlights frequency components in a logarithmic scale for drone identification. They demonstrated that Mel-spectrograms capture the unique acoustic characteristics of different drones, enabling more robust identification.

In recent years, deep learning techniques have emerged as powerful tools for pattern recognition and classification tasks. Transfer learning, in particular, has gained attention due to its ability to leverage pre-trained models on large datasets and adapt them to specific tasks with limited labeled data. One widely used deep learning model is the YAMNet neural network, originally designed for audio event recognition [25]. Table 1 determines the pros and cons of the various classification methods for unmanned aerial vehicles (UAVs).

## 3. Data characteristics

DroneRF consists of 227 segments by recording the signals of AR, Bebop, and Phantom drones as well as the background at low and high frequencies. All signals of drones are recorded from four different modes including on, hovering, flying, and video recording but Phantom drone is recorded from only one mode. For both low and high frequency, the segments are 41 of RF background activities and 21 of RF drone's mode. However, the AR drone's final mode only has 18 segments [26].

Fig. 1 illustrates all experiments classification performed on the dataset. In the initial experiment, all datasets can be considered as a 2-Class such as a drone and non-drone. The three types of drones can also be categorized into 3 Classes, namely AR, Bebop, and Phantom in the second classification experiment. Finally, all modes for every drone can be classified into 6-Class.

## 4. Preliminaries

### 4.1. YAMNet: audio neural network model

YAMNet (Yet Another Mobile Network) is a pre-trained neural network designed for audio waveform analysis [27]. It comprises a

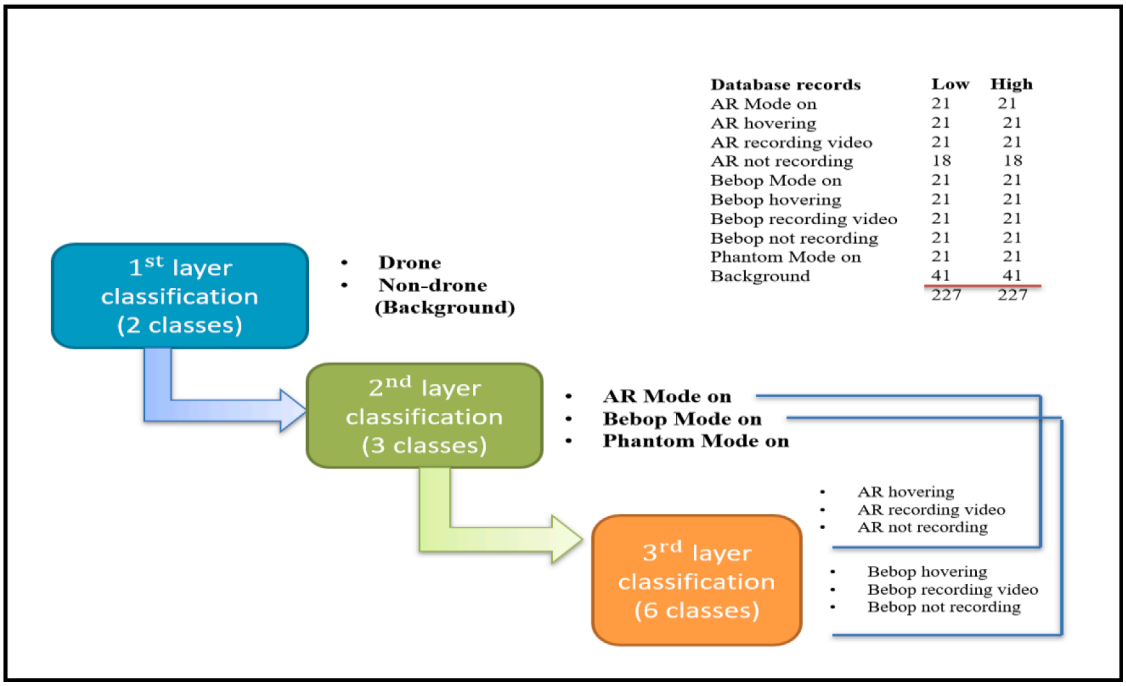



**Fig. 1.** Drone RF distribution and their categories.

total of 86 layers, which can be broken down as follows: 27 layers for convolutional operations (14 for standard convolution and 13 for depth-wise convolution), 27 layers for batch normalization and ReLU activation, 1 average pooling layer, 1 fully connected (FC) layer, 1 softmax layer, and a final classification layer [28].

The initial layer in YAMNet is the convolutional layer, responsible for extracting meaningful features from the input audio waveform. It is crucial to ensure that the data is properly normalized to avoid training difficulties and to enhance the network's learning speed. To achieve this, batch normalization layers are employed after each convolutional layer. These normalization layers not only accelerate the training process but also simplify the learning dynamics. Following the normalization layer, a rectified linear unit (ReLU) activation function is applied to prevent the network's computational complexity from growing exponentially.

During the classification phase, an FC layer and a softmax classifier are utilized. The FC layer consolidates information from each neuron in the preceding layer with every neuron in the subsequent layer. This enables comprehensive analysis of the entire input, facilitating informed decision-making.

By leveraging this architecture, YAMNet can effectively analyze audio waveforms and classify them based on learned patterns. Fig. 2 describes the YAMNet architecture.

### 4.2. Mel spectrogram of radio frequency

Spectrograms play a crucial role in the analysis and processing of sound signals, providing valuable insights into the frequency energy distribution over time. By establishing a comprehensive relationship between the time and frequency domains of sound signals, spectrograms maximize the extraction of sound feature information, thus contributing to the generation of fingerprinting features [29]. Spectrograms offer a wealth of information on the frequency scale. However, utilizing this information directly as audio signatures poses challenges due to the high resolution that may hinder distinct separation. To overcome this limitation, many audio fingerprinting algorithms adopt bandpass filter banks. These filter banks filter the audio signal, effectively reducing the information's resolution and improving the robustness of audio signatures in the form of more distinguishable spectrograms [30].

Upon examining the Mel spectrogram images, it becomes evident that the duplicated columns within the spectrogram remain unaltered even after subjecting the audio to various post-processing operations. This observation underscores the effectiveness of Mel spectrogram images as a robust approach for extracting features from audio. The significance of the Mel spectrogram stems from its adherence to the Mel scale, which closely emulates human auditory perception. As a result, it provides a reliable representation of the frequencies typically perceived by humans.

The Mel scale is a perceptual scale that captures the relative distances between pitches as perceived by listeners. In contrast to standard frequency measurements, the Mel scale establishes a reference point by assigning a perceptual pitch of 1000 Mels to a 1000 Hz tone, which is 40 dB above the listener's threshold. This calibration allows for a more accurate representation of pitch perception based on human auditory capabilities. The process of generating the fingerprint spectrogram, as depicted in Fig. 3, involves framing, windowing, and applying the discrete Fourier transform (DFT) [31]. To begin with, a segment of a 2-second noise waveform is extracted, and the Hamming window is selected to ensure smoothness and minimize signal distortion at both ends of the segment. The segment is then subjected to framing, where the choice of frame length is critical for preserving the accuracy of feature extraction. If the frame length is too long, it may adversely affect the accuracy, while a frame length that is too short may result in the loss of useful features. Considering that transformer noise exhibits greater stability compared to the human voice, a slightly longer frame length can be used to ensure the integrity of audio signal characteristics. Additionally, to ensure a smooth transition between frames, an overlap

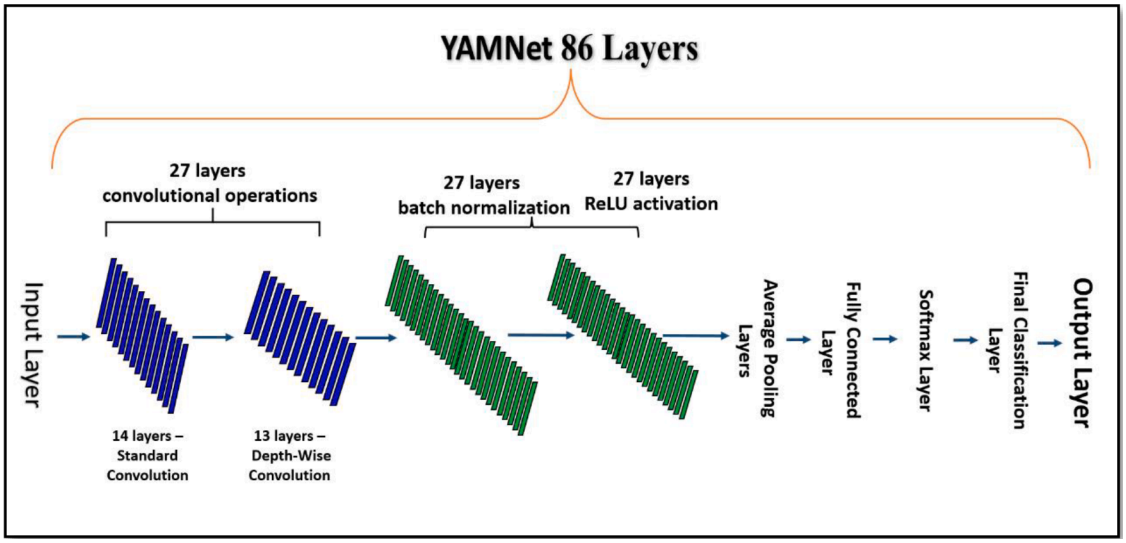



**Fig. 2.** YAMNet architecture.

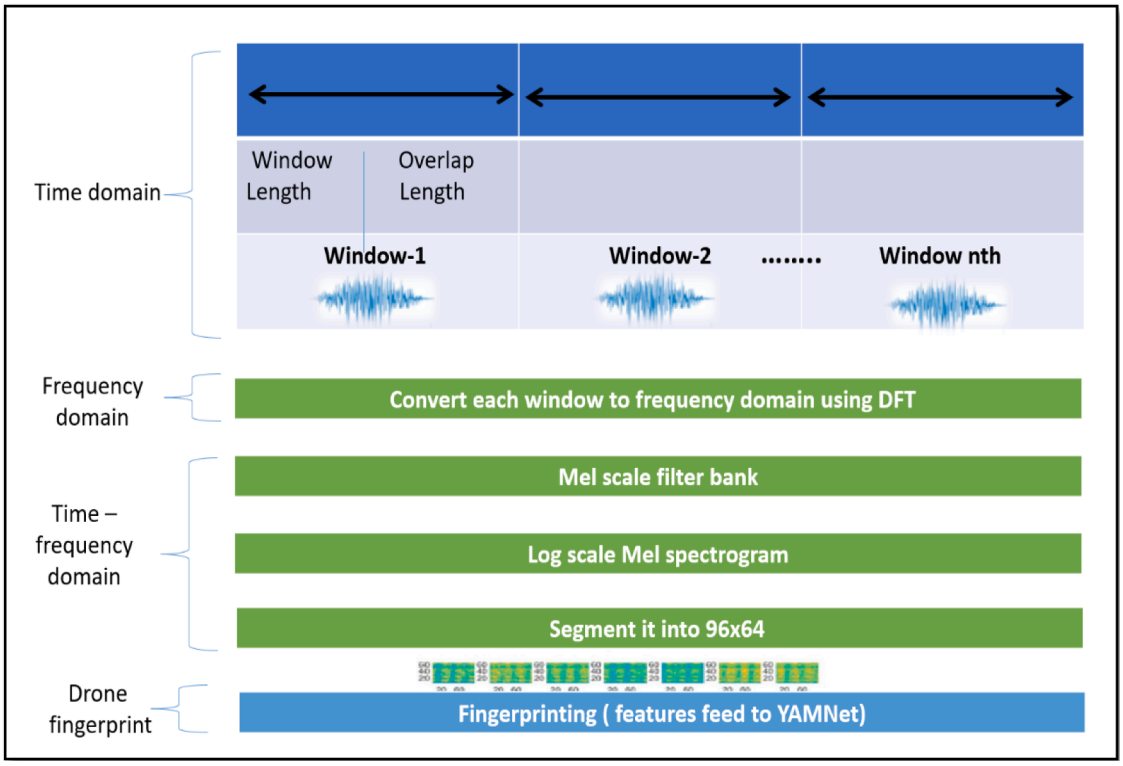



**Fig. 3.** Mel-Spectrogram: Time domain to Frequency domain to Time-frequency domain transformation.

rate of 95% of the window length is employed, with a hop size of 10 ms.

After framing, it is important to address the issue of spectrum leakage that occurs if the DFT is directly applied to the framed data. Thus, each frame undergoes windowing before further processing. Moreover, the audio signal is decomposed into separate frequency

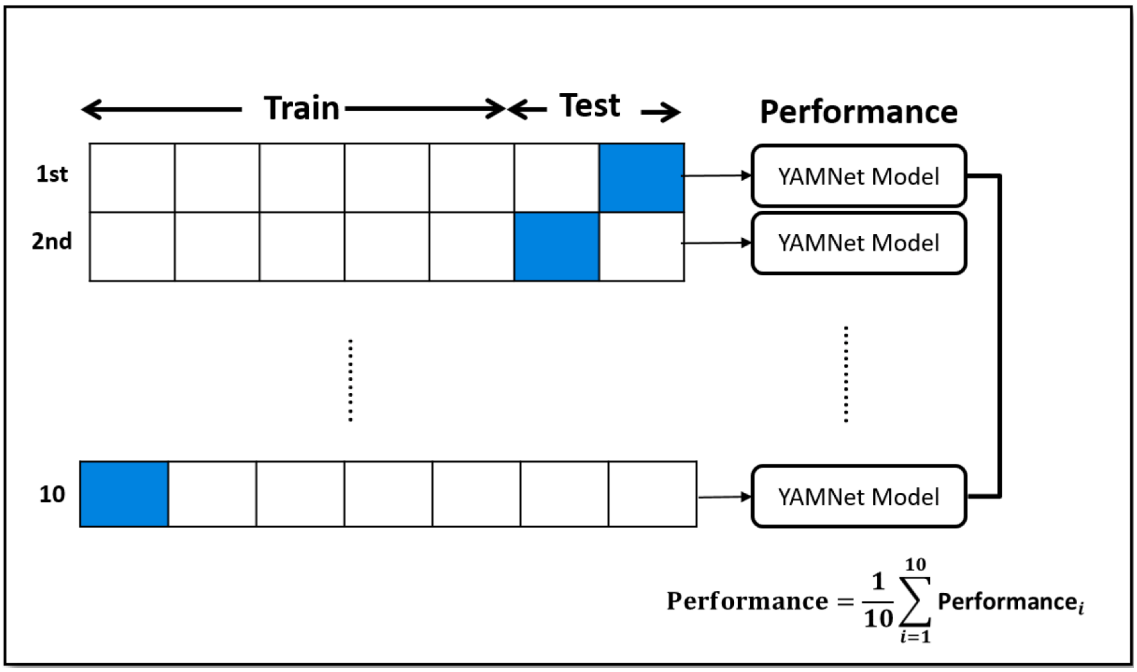



**Fig. 4.** The mechanism of the cross-validation strategy.

bands using a Mel filter bank, which operates in the Mel frequency scale to mimic the nonlinear characteristics of human auditory perception [32]. This decomposition into distinct frequency bands enables a more effective representation of audio features.

### 4.3. Cross-validation

Cross-validation is a widely employed resampling technique utilized to assess the model's performance [33]. One commonly used approach is K-fold cross-validation, where the input data is divided into two subsets: the training set and the test set. Initially, the data is partitioned into k subsets, with K-1 groups allocated for training the model, while the remaining group is held out for testing and evaluating the model's performance [34]. This iterative process ensures that each subset serves as both the training and test set, allowing for a comprehensive evaluation of the model's generalization capability.

In the next iteration, select a different subset for the test data and retrain the model. In the end, the training process is repeated k-times, and calculate the best evaluation by taking the average of all individual evaluations. Interchanging between the training and test sets in K-fold cross-validation increases the effectiveness of the model and avoids overfitting. Fig. 4 describes the mechanism of the cross-validation strategy.

## 5. The proposed Mel spectrogram-based fingerprinting drone classification model

The proposed model for drone classification based on Mel Spectrogram-based fingerprinting comprises three essential building phases. These phases are as follows:

1 RF drone dataset preparation phase: In the initial phase of the study, a series of processing steps are conducted to convert the radio frequency signals into audio data. Subsequently, the audio data is further transformed into Mel spectrogram images.
2 Spectrogram fingerprinting generation phase: The second phase focuses on generating drone fingerprints using spectrogram data. This phase involves a specific procedure designed to extract distinctive fingerprinting features for drones.
3 Classification phase: The final phase involves the classification and prediction of drone types and modes using the YAMNet deep learning framework. This phase utilizes the generated fingerprints to train the model and enable accurate classification.

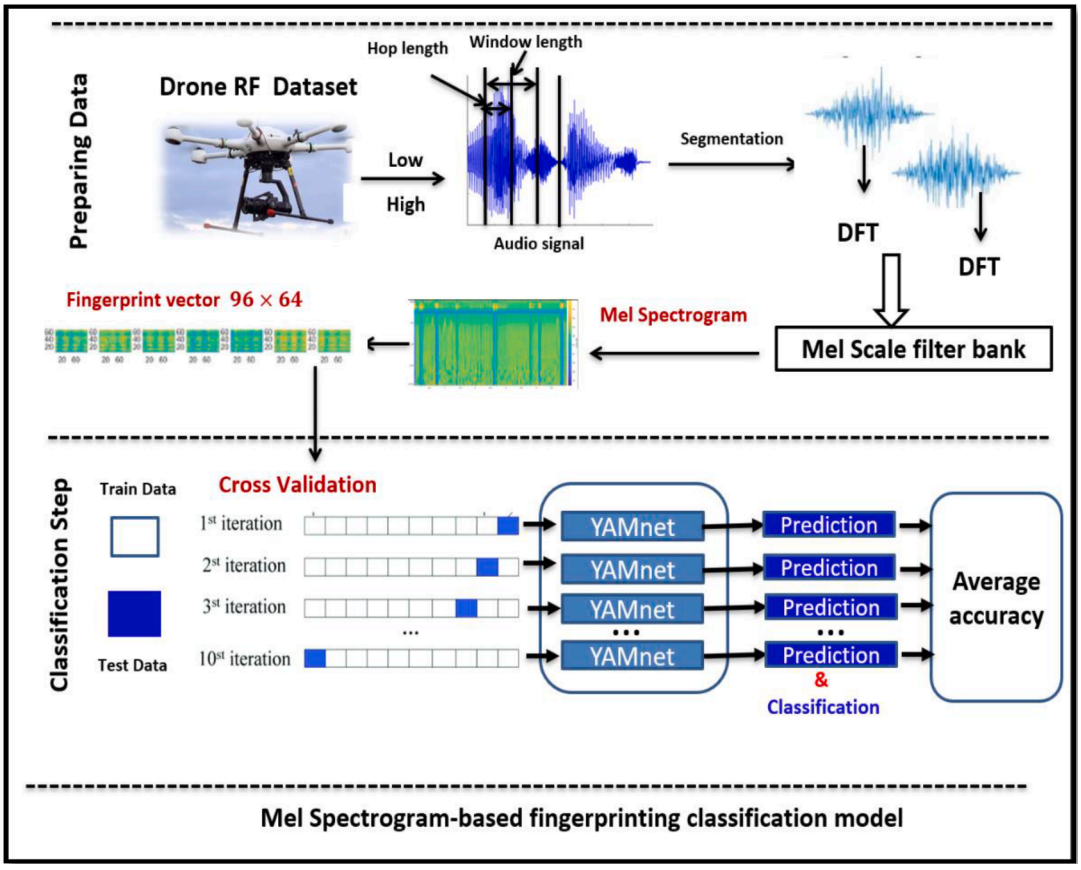



**Fig. 5.** The general architecture of the prediction and classification of the drone type and mode.

These three phases are elaborated in detail in the subsequent section, encompassing the steps involved and the distinctive characteristics associated with each phase. The overall architecture of the proposed model is visually depicted in Fig. 5, providing an overview of the system's structure and flow.

### 5.1. Preparing the RF drone data sets the phase

The drone's RF, which is stored in both low-band and high-band RF signals of the data set, is distributed across the frequency spectrum. So, the properties of drone signals must be retrieved more effectively in the frequency domain.

Initially, the first step is to convert the RF file to an audio signal. Then, the audio signals are divided into segments using a 25-ms periodic Hann window (400 samples) along with a 95% overlap of the window's length, a 10-ms hop, and a Short-Time Fourier Transform (STFT). STFT is a common technique used for analyzing the frequency content of an audio signal over time. After that, the discrete Fourier transform (DFT) is applied individually on each segment to convert each window from time to frequency domain. DFT calculates as the following:

$$y_i = \sum_{n=1}^{N} x_i\ (n)\ \exp\left(\frac{-j2\pi m(n-1)}{M}\right) \tag{1}$$

where $x_i$ is the $i^{th}$ RF segment, $y_i$ is the ranges of the $i^{th}$ segments, $n$, and $m$ are the time and frequency domain, $N$ is the total number of time samples in the RF segment.

### 5.2. Spectrogram fingerprinting generation phase

Spectrogram-based n-dimensional vectors per frame are generated from all the audio references and stored in reference fingerprints database. This part describes the main procedure to generate the drone fingerprint using the radio frequency signals.

The main procedure of the generation of the drone fingerprinting is described on Algorithm 1 and Fig. 6 depicts these steps.

### 5.3. Drone classification phase: cross-validation-based

To develop an effective identification and classification of drones, the YAMNet neural network is utilized, which consists of 86 layers. The first layer in YAMNet is the convolution layer, which is responsible for identifying meaningful features from the input image. After each convolutional layer, a batch normalization layer is applied to speed up the training and simplify the learning process.

**Algorithm 1**
Mel Spectrogram based drone fingerprint generation.

**Input:** CSV DroneRF files
**Output:** Drones Fingerprint
**1. For** each CSV file **Do**
2. Convert CSV DroneRF into audio files.
3. Divide the audio signals into frames of 30 ms duration with the Hamming window
4. Apply discrete Fourier transform (DFT) to all the subframes
**5. End**
**6. For** each audio **Do**
7. Calculate the Mel scale filter bank using the following equation:

$$H_k(m) = \begin{cases} 0 \text{ for } m < f(k-1) \\ \dfrac{m - f(k-1)}{f(k) - f(k-1)} \text{ for } f(k-1) \le m \le f(k) \\ \dfrac{f(k+1) - m}{f(k+1) - f(k)} \text{ for } f(k) < m \le f(k+1) \\ 0 \text{ for } m > f(k+1) \end{cases} \tag{2}$$

**Where** $H_k(m)$ is the triangular window function, $m$ is the number of filter, and $k$ ranges from 0 to $m - 1$.

8. Generation and presentation of spectrogram images
**Where**
the y-axis in spectrogram images corresponds to the Mel domain, while the x-axis represents the time dimension of the data
9. Divided the spectrogram matrix (images) into small windows $(w_i, h_j)$
**10. For** each window **Do**
11. compute the mean using the following form:

$$Means_{window} = \sum_{i=j=1}^{2} \frac{(w_i, h_j)}{4} \tag{3}$$

12. Multiply sum of means for each window to the spectrogram matrix.
13. Separate the horizontal and vertical slices of the spectrogram matrix
14. Concatenate the two slices to generate the feature fingerprint vector.
15. Store all generated fingerprints into database
**16. End**
**17. End**

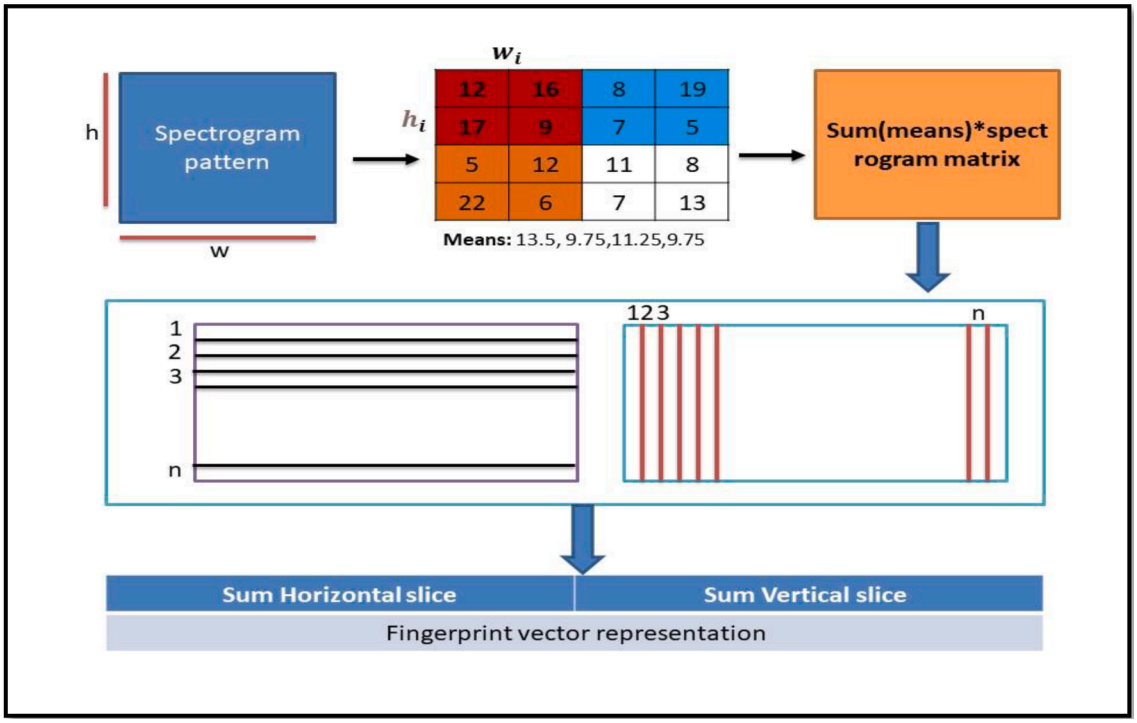



**Fig. 6.** Spectrogram fingerprinting generation.

Normalization is necessary because non-normalized data may cause training difficulties and slow down the network's rate of learning. Following the normalization layer, a ReLU layer is applied to prevent the computation in the neural network from growing at an exponential rate.

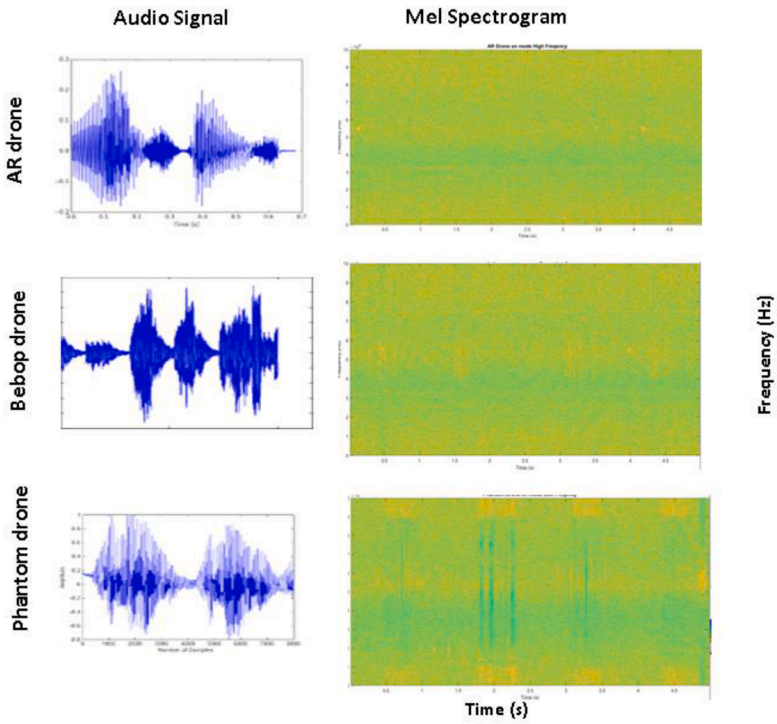



**Fig. 7.** Time-frequency representations of different drones.

In the classification phase, the fully connected layer of the CNN plays a crucial role. This layer facilitates the integration of information from all neurons in the preceding layer with those in the subsequent layer, allowing decisions to be made based on the entirety of the image. To achieve the final classification, the output of the fully connected layer is passed through a softmax classifier. The incorporation of the YAMNet neural network and transfer learning techniques enhanced the model's classification capabilities, enabling accurate identification of different drone types and modes.

During the training and testing of the YAMNet neural network, a 10-fold cross-validation strategy is employed. This strategy ensures that all the available data is utilized during the testing stage, and the performance of the model is evaluated based on the average performance over the 10-fold validation. This approach provides a comprehensive assessment of the model's overall performance.

## 6. Experimental results and the evaluation

In this section, the evaluation metrics and results of the experiments are presented. The experiments utilize a GPU specifically optimized for the MATLAB R2022b software package. For all experimental procedures, a computer system equipped with a core i7 processor and 16 GB of RAM is employed, with the GPU handling all experiments.

The first step in our experiment is the preprocessing step. The data pre-processing step includes resampling the audio signal and generating an array of Mel spectrograms as shown in Fig. 7. The first step is to convert the RF file for various drones to an audio signal. Subsequently, the audio signals are divided into L overlapping segments and resampled to 16 kHz. To generate the Mel spectrogram, a 25-ms periodic Hann window (400 samples) is used, along with a 95% overlap of the window's length and a 10-ms hop. The final result is the time-frequency representation of the audio signal.

The creation and presentation of spectrogram images are part of the procedure. The Mel domain is represented by the y-axis in these images, while the data's time dimension is represented by the x-axis. Then, the spectrogram pictures were partitioned into little windows. For each window, find the mean. Finally, separate the horizontal and vertical slices of the spectrogram grid as shown in Fig. 8.

To use the Mel spectrogram as input for YAMNet, a $96 \times 64 \times 1$ L array is employed to represent the input audio. Here, 96 represents the number of Mel spectrograms, 64 represents the number of Mel bands, and L is determined by the length of the input audio. The resulting input layer for YAMNet is a $96 \times 64$ array of Mel spectrograms. Initially, YAMNet is pre-trained with more than 1 Million AudioSet-YouTube corpora to classify 521 sound occasion classes. The evaluation set and training data details are listed in Table 2.

For parameter optimization, a 10-fold cross-validation performance has been considered. For the purpose of classifying drones, all optimized parameters are selected and compiled in Table 3. With a learning rate of 0.0001, the Adam optimizer is one of the hyperparameters used to update biases and weights to reduce the error. Cross entropy is used as a loss function to calculate the difference between the ground and predicted values. In addition, epoch for training process is set to 100.

Different metrics are utilized to measure the performance of experimental results such as accuracy, precision, F-Measure, and recall. The definitions of the metrics are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1} - \text{Score} = \frac{2TP}{2TP + FP + FN} \tag{6}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{7}$$

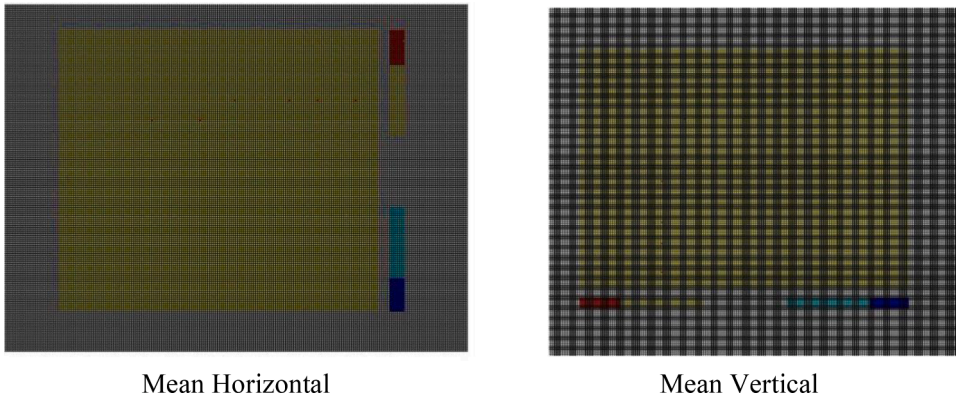



**Fig. 8.** The horizontal and vertical slices of the spectrogram matrix.

**Table 2**
YAMNet characteristic.

| Parameter | Value |
|---|---|
| No. of layer | 86 |
| Parameter | 3.75 Million |
| Size of input | 96 × 64 × 1 |
| Training data size | 1789,621 audio segments |
| Class | 521 audio class |

**Table 3**
Details of fine-tuning parameters of optimizer.

| Parameter | Value |
|---|---|
| Optimizer | Adam |
| Learning Rate | 0.0001 |
| Loss Function | Cross entropy |
| Metrics | accuracy |
| Batch Size | 15 |
| Epochs | 100 |

$$Specificity \;=\; TNR = \frac{TN}{TN + FP} \tag{8}$$

Where $FP,\ FN,\ TP$ *and* $TN$ is the False Positive, False Negative, True Positive, and True Negative, respectively.

### 6.1. Evaluation

The proposed approach was evaluated through three experiments. The first experiment involved detecting the presence or absence of a drone. Table 4 represents the classification performance of the first layer classification during 10 iterations and over different batch size: 100, 50, 25 and 15. At every batch size, the training progress is observed and calculates the average for all individual evaluation to calculate the best evaluation. This method ensures that the training is completed as quickly as possible and stops the model from over-fitting. Results show that the model obtains the best result at 15 batch size. The results showed a 99.6% accuracy, 100% sensitivity, 99.23% specificity, and 99.6% F1-score. In addition, the extracted features are analyzed using t-SNE [35], as depicted in Fig. 9. The variety type blue and red addresses the two classes. It gives that our model creates a superior partition of classes

In the second classification experiment, the proposed approach was evaluated for its ability to identify different types of drones (AR drone, Phantom drone, and Bebop drone). The effect of 10 cross validation in generating input has been studied further by varying batch size (25, 50, and 100). The results are organized in Table 5. At 50 batch size, the accuracy average gets 91.92% while it is achieved 89.04% at 100 batch size. According to the study, the accuracy is 95.25% and the model performance improves significantly with batch size of 25.

Table 6 shows the results that were obtained for each of the three classes based on various performance criteria. The results show that, the batch size at 15, the accuracy provides the best evaluation metrics. Following 15 batch size is 25, which has an approximate accuracy of 95.25%. As a result, this is kept as the best value. Additionally, the extracted features are analyzed by t-SNE, as shown in Fig. 10. The three classes are covered by the blue, red, and yellow varieties. It demonstrates that our model produces a superior class partition.

In the third grouping test, the proposed approach was assessed for its capacity to distinguish drone modes. By varying batch size (25, 50, and 100), the effect of 10 cross validation on generating input has been investigated further. The outcomes are coordinated in Table 7. The average accuracy is 93.01% at 50 batch sizes, while it is 85.51% at 100 batch sizes. The study found that with a batch size of 25, the model performs significantly better and has an accuracy of 94.49%.

For the final classification layer, the proposed approach was evaluated specifically for identifying the mode of each type with different metrics. As the batch size of 15, the best evaluation metrics are provided by accuracy. Fig. 11 shows an accuracy of 96.09%, sensitivity of 96.27%, specificity of 98.26%, and F1-score of 59.84% for AR drone modes. Fig. 12 show the results showed an accuracy of 97.56%, sensitivity of 97.56%, specificity of 98%, and F1-score of 97.7% for bebop drone at 15 batch size. Overall, the experimental results demonstrate that the proposed approach is effective in detecting the presence of drones and accurately identifying their types and modes based on RF signals.

The comparison of our proposed model to other fingerprinting algorithms is shown in Fig. 13. The discrete Fourier transform is used to extract the radio frequency signal's frequency spectrum as the UAV's fingerprint in [21], with an identification accuracy of 845.5%. In [22], physical characteristics like body shifting and vibration are chosen as UAV fingerprints. The identification accuracy of this scheme can reach up to 89.7%. Short-time Fourier transform performs in [23] to acquire the time-recurrence range and then the PCA algorithm used to extract fingerprints. At the end, SVM is used for identification, with 94.7% accuracy.

**Table 4**
Evaluation metrics (in mean) over different batch size in first classification.

| Batch Size | Accuracy% | Sensitivity% | Specificity% | Precision% | F1-score% |
|---|---|---|---|---|---|
| 100 | 85.339 | 92.088 | 86.689 | 81.987 | 84.08 |
| 50 | 95.615 | 95.646 | 96.667 | 96.154 | 95.529 |
| 25 | 99.215 | 99.286 | 99.231 | 99.231 | 99.23 |
| 15 | 99.6 | 100 | 99.23 | 99.23 | 99.6 |

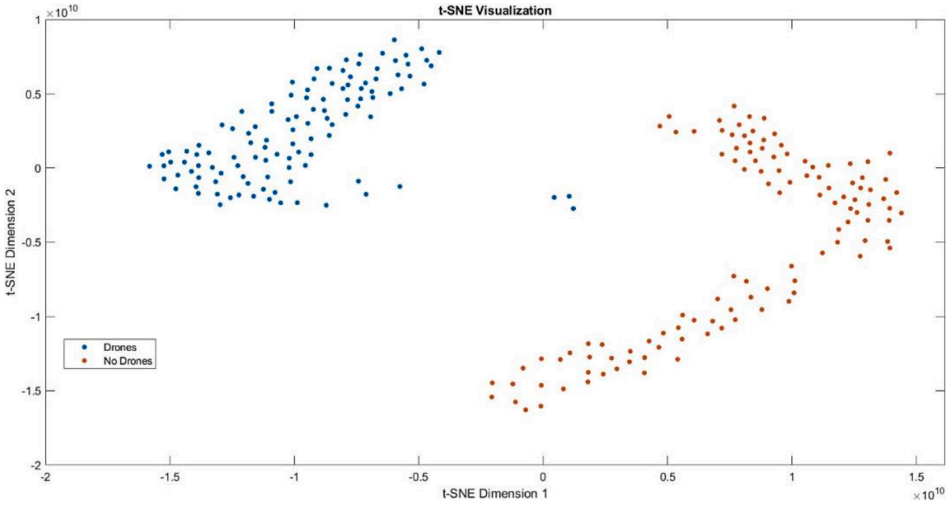



**Fig. 9.** The t-SNE representation of feature for 2-class.

**Table 5**
Performance evaluation for the second classification layer at different batch size.

| Cross Validation | Accuracy% 25 | 50 | 100 | Sensitivity% 25 | 50 | 100 | F1-score% 25 | 50 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | 83.33 | 91.67 | 100 | 85.00 | 93.33 | 100 | 83.20 | 91.53 |
| 2 | 100 | 92.31 | 92.31 | 100 | 93.33 | 93.33 | 100 | 91.53 | 91.53 |
| 3 | 100 | 100 | 84.62 | 100 | 100 | 86.11 | 100 | 100 | 83.87 |
| 4 | 100 | 92.31 | 76.92 | 100 | 94.44 | 80.00 | 100 | 92.21 | 76.85 |
| 5 | 92.3 | 92.31 | 84.62 | 93.3 | 94.44 | 88.89 | 92.59 | 92.21 | 85.00 |
| 6 | 84.6 | 92.31 | 100 | 87.7 | 93.33 | 100 | 82.15 | 91.53 | 100 |
| 7 | 92.3 | 100 | 76.92 | 93.3 | 100 | 87.50 | 91.53 | 100 | 76.43 |
| 8 | 91.7 | 75.00 | 83.33 | 93.3 | 82.22 | 88.89 | 91.53 | 69.63 | 82.22 |
| 9 | 91.6 | 91.67 | 100 | 93.3 | 93.33 | 100 | 91.53 | 91.53 | 100 |
| 10 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| **average** | **95.25** | 91.92 | 89.04 | **96.11** | 93.61 | 91.81 | **94.93** | 91.184 | 88.743 |

**Table 6**
The performance (in mean) over different batch size in the second classification.

| Batch Size | Accuracy% | Sensitivity% | Specificity% | Precision% | F1-score% |
|---|---|---|---|---|---|
| 100 | 89.04 | 91.81 | 95.24 | 89 | 88.74 |
| 50 | 91.92 | 93.61 | 96.6 | 91.67 | 91.18 |
| 25 | 95.25 | 96.11 | 97.95 | 95.17 | 94.93 |
| 15 | **96.1** | **96.7** | **98.26** | **96** | **95.87** |

## 7. Conclusion and future work

In this paper, a new approach for drone identification and classification based on Mel spectrogram-based RF fingerprints using a pre-trained YAMNet neural network has been proposed. Our model successfully converted radio frequency signals into audio signals and extracted Mel spectrograms as unique fingerprints for drone identification. The YAMNet neural network, adapted through transfer learning, demonstrated excellent performance in classifying drones and distinguishing between drone types and modes.

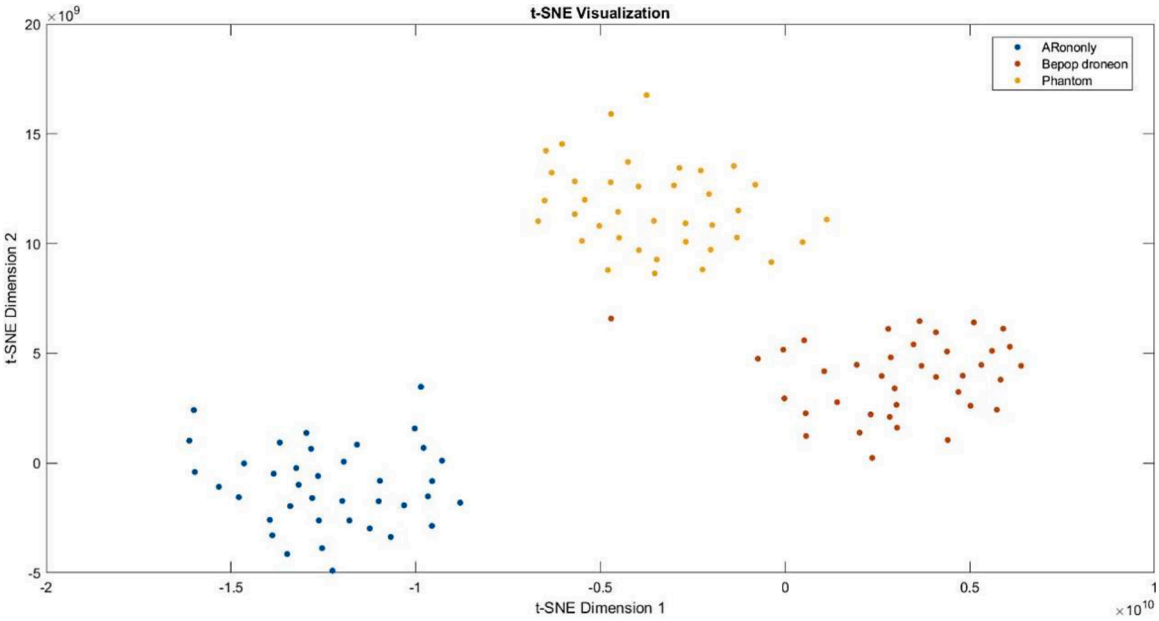



**Fig. 10.** The t-SNE representation of feature for 3-class.

**Table 7**
Performance evaluation for the third classification layer at different batch size.

| Cross Validation | Accuracy% 25 | 50 | 100 | Sensitivity% 25 | 50 | 100 | F1-score% 25 | 50 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | 100 | 83.33 | 100 | 100 | 85.00 | 100 | 100 | 83.20 |
| 2 | 100 | 92.31 | 100 | 100 | 94.44 | 100 | 100 | 92.21 | 100 |
| 3 | 92.31 | 76.92 | 92.31 | 93.33 | 82.22 | 94.44 | 91.53 | 75.34 | 92.21 |
| 4 | 100 | 84.62 | 92.31 | 100 | 85.00 | 93.33 | 100 | 85.00 | 92.59 |
| 5 | 84.62 | 92.31 | 92.31 | 90.48 | 94.44 | 94.44 | 84.92 | 92.21 | 92.21 |
| 6 | 92.31 | 100 | 69.23 | 93.33 | 100 | 77.38 | 92.59 | 100 | 69.44 |
| 7 | 92.31 | 92.31 | 92.31 | 93.33 | 93.33 | 94.44 | 92.59 | 92.59 | 92.21 |
| 8 | 100 | 91.67 | 41.67 | 100 | 93.33 | 57.78 | 100 | 91.53 | 41.48 |
| 9 | 100 | 100 | 91.67 | 100 | 100 | 93.33 | 100 | 100 | 91.53 |
| 10 | 83.33 | 100 | 100 | 83.33 | 100 | 100 | 83.33 | 100 | 100 |
| Average | **94.49** | 93.01 | 85.51 | **95.38** | 94.28 | 89 | **94.5** | 92.9 | 85.49 |

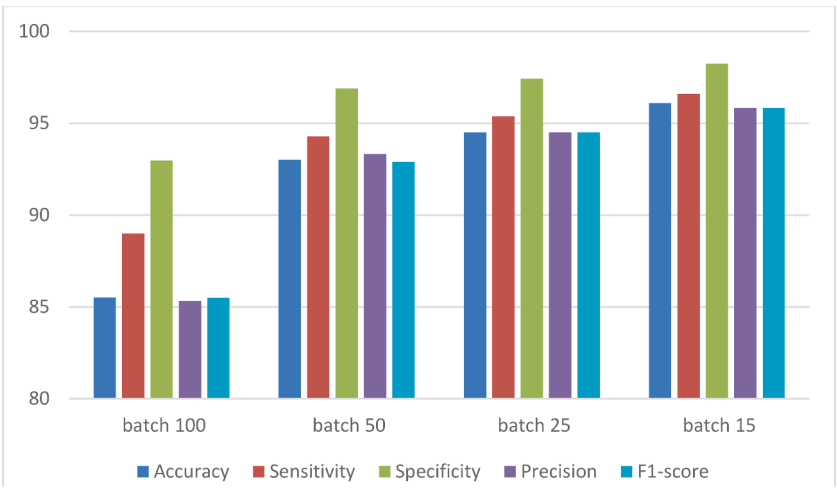



**Fig. 11.** Model evaluation metrics (in mean) over different batch size for AR modes.

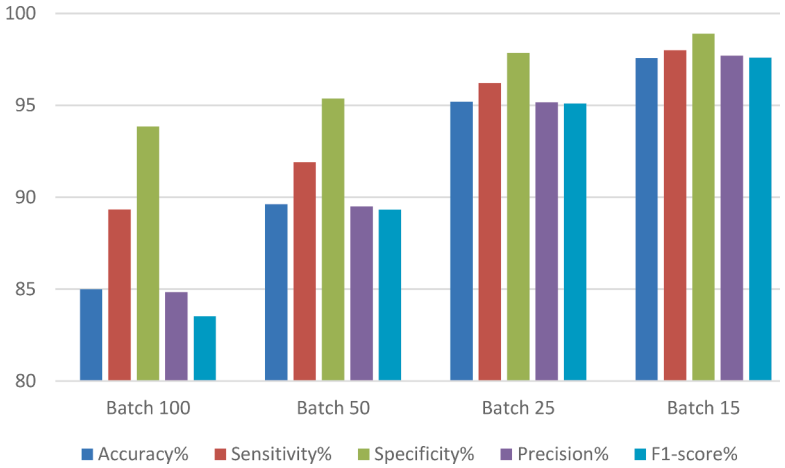



**Fig. 12.** Model evaluation metrics (in mean) over different batch size for Bebop modes.

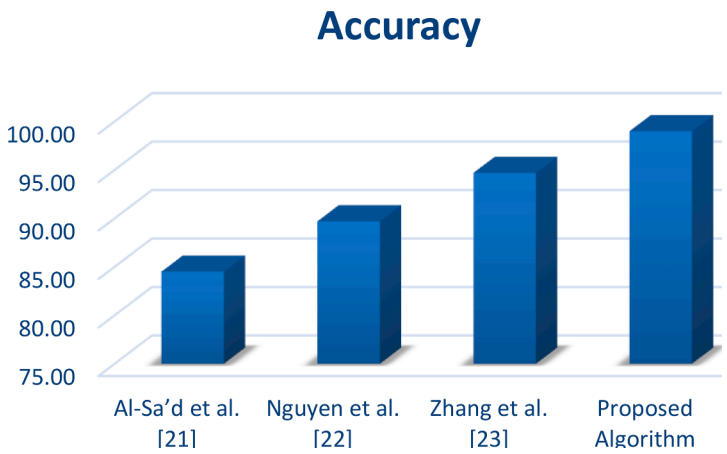



**Fig. 13.** Comparison between fingerprints in different algorithms.

Through extensive experimentation and evaluation, our model achieved high accuracy levels. In the first classification layer, we achieved a remarkable accuracy of 99.6% in distinguishing between drones and non-drones. The second classification layer achieved 96.9% accuracy in identifying drone types from three classes, including AR Drone, Bebop Drone, and Phantom Drone. Additionally, the third classification layer achieved an accuracy of 96–97% in identifying the mode of each drone type.

The Mel spectrogram-based RF fingerprinting approach proved effective in capturing important audio feature information for drone identification. The utilization of the Mel frequency scale closely mimicked human perception, providing a robust representation of the frequencies typically heard by humans.

Future work in this field could involve expanding the dataset to include a wider range of drone types and modes, further enhancing the model's classification accuracy. Additionally, investigating the impact of different preprocessing techniques and optimization methods on the performance of the model would be valuable. Furthermore, exploring the potential integration of other audio analysis techniques, such as deep learning architectures specifically designed for audio signals, could contribute to advancing the field of drone identification and classification.

This research provides valuable insights into the application of Mel spectrogram-based RF fingerprints and pre-trained neural networks for drone identification and classification, paving the way for enhanced drone surveillance and security systems in various domains.

## Declaration of Competing Interest

There is no conflict of Interest.

## Data availability

Data will be made available on request.

## References


[1] Maria Stoyanova, et al., A survey on the internet of things (IoT) forensics: challenges, approaches, and open issues, IEEE Commun. Surveys & Tutorials 22 (2) (2020) 1191–1221.
[2] G. Macrina, L. Di Puglia Pugliese, F. Guerriero, G. Laporte, Drone-aided routing: a literature review, Transp. Res. Part C: Emerg. Technol. 120 (Nov. 2020), 102762, https://doi.org/10.1016/j.trc.2020.102762.
[3] S.A.H. Mohsan, M.A. Khan, F. Noor, I. Ullah, M.H. Alsharif, Towards the unmanned aerial vehicles (UAVs): a comprehensive review, Drones 6 (6) (Jun. 2022) 147, https://doi.org/10.3390/drones6060147.
[4] J.P. Yaacoub, H. Noura, O. Salman, A. Chehab, Security analysis of drones systems: attacks, limitations, and recommendations, Internet of Things 11 (Sep. 2020), 100218, https://doi.org/10.1016/j.iot.2020.100218.
[5] J. Pyrgies, The UAVs threat to airport security: risk analysis and mitigation, J. Airl. Airpt. Manag. 9 (2) (Oct. 2019) 63, https://doi.org/10.3926/jairm.127.
[6] Nicolás Molina-Padrón, et al., Monitoring in near-real time for amateur UAVs using the AIS, IEEE Access 8 (2020) 33380–33390.
[7] Yu Wu, Kin Huat Low, An adaptive path replanning method for coordinated operations of drone in dynamic urban environments, IEEE Syst. J. 15 (3) (2020) 4600–4611.
[8] Tiago M. Fernández-Caramés, et al., Towards an autonomous industry 4.0 warehouse: a UAV and blockchain-based system for inventory and traceability applications in big data-driven supply chain management, Sensors 19 (10) (2019) 2394.
[9] J. Gong, X. Xu, Y. Lei, Unsupervised specific emitter identification method using radio-frequency fingerprint embedded infoGAN, IEEE Trans. Inf. Forensics Secur. 15 (2020) 2898–2913.
[10] H.R. Su, K.Y. Chen, W.J. Wong, S.H. Lai, A deep learning approach towards pore extraction for high-resolution fingerprint recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), New Orleans, LA, USA 5–9, March 2017, pp. 2057–2061.
[11] D. Roy, T. Mukherjee, M. Chatterjee, E. Blasch, E.RFAL Pasiliao, Adversarial learning for RF transmitter identification and classification, IEEE Trans. Cogn. Commun. Netw. 6 (2020) 783–801.
[12] Samiur Rahman, Duncan A. Robertson, Classification of drones and birds using convolutional neural networks applied to radar micro-Doppler spectrogram images, IET Radar, Sonar & Navigation 14 (5) (2020) 653–661.
[13] Se-Won Yoon, et al., Efficient classification of birds and drones considering real observation scenarios using FMCW radar, J. Electromagnetic Eng. Sci. 21 (4) (2021) 270–281.
[14] Carl Chalmers, et al., Video analysis for the detection of animals using convolutional neural networks and consumer-grade drones, J. Unmanned Vehicle Syst. 9 (2) (2021) 112–127.
[15] C. Aker, S. Kalkan, Using deep networks for drone detection, in: Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 2017, 29 August–1 September.
[16] A. Schumann, L. Sommer, J. Klatte, T. Schuchert, J. Beyerer, Deep cross-domain flying object classification for robust UAV detection, in: Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 2017, 29 August–1 September.
[17] B. Taha, A. Shoufan, Machine learning-based drone detection and classification: state-of-the-art in research, IEEE Access 7 (2019) 138669–138682.
[18] M.Z. Anwar, Z. Kaleem, A. Jamalipour, Machine learning inspired sound-based amateur drone detection for public safety applications, IEEE Trans. Veh. Technol. (2019) 2526–2534.
[19] Z. Shi, X. Chang, C. Yang, Z. Wu, J. Wu, An acoustic-based surveillance system for amateur drones detection and localization, IEEE Trans. Veh. Technol. 69 (2020) 2731–2739.
[20] Wei Nie, et al., UAV detection and identification based on WiFi signal and RF fingerprint, IEEE Sens. J. 21 (12) (2021) 13540–13550.
[21] M.F. Al-Sa'd, A. Al-Ali, A. Mohamed, T. Khattab, A. Erbad, RFbased drone detection and identification using deep learning approaches: an initiative towards a large open source drone database, Future Gener. Comput. Syst. 100 (Nov. 2019) 86–97.
[22] P. Nguyen, H. Truong, M. Ravindranathan, A. Nguyen, R. Han, T. Vu, Matthan: drone presence detection by identifying physical signatures in the drone's RF communication, in: *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, New York, NY, USA, Jun. 2017, pp. 211–224.
[23] P. Zhang, L. Yang, G. Chen, G. Li, Classification of drones based on micro-Doppler signatures with dual-band radar sensors, in: Proc. Prog. Electromagn. Res. Symp. - Fall (PIERS-FALL), Nov. 2017, pp. 638–643.
[24] Martins Ezuma, et al., Micro-UAV detection and classification from RF fingerprints using machine learning techniques, in: 2019 IEEE Aerospace Conference. IEEE, 2019.
[25] Alberto Tena, Francesc Claria, Francesc Solsona, Automated detection of COVID-19 cough, Biomed. Signal Process. Control 71 (2022), 103175.
[26] Mohammad Al-Sa'd, Mhd Saria Allahham, Amr Mohamed, Abdulla Al-Ali, Tamer Khattab, Aiman Erbad, DroneRF dataset: a dataset of drones for RF-based detection, classification, and identification, Mendeley Data, v1 (2019). 10.17632/f4c2b4n755.1.
[27] Ievgeniia Kuzminykh, et al., Audio interval retrieval using convolutional neural networks, in: Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 20th International Conference, NEW2AN 2020, and 13th Conference, ruSMART 2020, St. Petersburg, Russia, August 26–28, 2020, Proceedings, Part I 20, Springer International Publishing, 2020.
[28] Félix Gontier, Romain Serizel, Christophe Cerisara, Automated audio captioning by fine-tuning bart with audioset tags, Detection and Classification of Acoustic Scenes and Events-DCASE (2021).
[29] Z. Abduh, E.A. Nehary, M.A. Wahed, Y.M. Kadah, Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers, Biomed. Signal Process. Control 57 (2020), 101788.
[30] Nebras Sobahi, et al., Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds, Biocybernetics and Biomed. Eng.42 3 (2022) 1066–1080.
[31] Arnab Maity, Akanksha Pathak, Goutam Saha, Transfer learning based heart valve disease classification from Phonocardiogram signal, Biomed. Signal Process. Control 85 (2023), 104805.
[32] Sanjana Patil, Kiran Wani, Gear fault detection using noise analysis and machine learning algorithm with YAMNet pretrained network, Mater. Today: Proceed. 72 (2023) 1322–1327.
[33] A. Ramezan, Timothy A. Warner Christopher, Aaron E. Maxwell, Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification, Remote Sens. (Basel) 11 (2) (2019) 185.
[34] Jerzy Wieczorek, Cole Guerin, Thomas McMahon, K-fold cross-validation for complex sample surveys, Stat 11 (1) (2022) e454.
[35] Laurens Van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Machine Learn. Res. 9 (2008) 11.