ISSN: 2182-2069 / E-ISSN: 2182-2077

Real Time Drone Detection Based on Acoustics Using Hybrid Deep Learning Models

Shahad W. Jasim[1*], and Saad S. Hreshee[2]

[1*]Department of Electrical Engineering, University of Babylon, Babylon, Iraq.
eng571.shahad.waleed@student.uobabylon.edu.iq, https://orcid.org/0009-0002-3815-3136

[2]Department of Electrical Engineering, University of Babylon, Babylon, Iraq.
eng.saad.saffah@uobabylon.edu.iq, https://orcid.org/0000-0002-7547-1443

Received: February 21, 2025; Revised: March 31, 2025; Accepted: May 09, 2025; Published: May 30, 2025

Abstract

The rapid advancement of unmanned aerial vehicle (UAV) technology has increased the accessibility and capabilities of drones, enabling various applications but also raising serious cybersecurity and public safety concerns. Drones can be misused for surveillance, smuggling, or sabotage—especially in environments where visual or radio frequency (RF)-based detection is limited. In such cases, acoustic detection offers a promising alternative by leveraging the unique sound signatures emitted by drones. This paper proposes a hybrid deep learning approach for drone detection and classification based on acoustic signals. The method combines Conformer-based architectures, which integrate self-attention mechanisms from Transformer models, with traditional deep learning components—Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Convolutional Recurrent Neural Networks (CRNN). These combinations aim to improve both spatial feature extraction and sequential modeling under real-world conditions. We developed and tested three hybrid models: (1) CNN + Conformer, which captures spatial features and long-term dependencies; (2) CRNN + Conformer, combining convolutional, recurrent, and attention-based features for enhanced contextual understanding; and (3) RNN + Conformer, focused on improving sequence learning in noisy environments. A custom dataset of 1,500 drone audio clips was collected using a parabolic microphone under varying environmental conditions. The recordings, which include two commercial drone models (DJI Phantom 4 Pro and DJI Mavic 2 Pro), were augmented to simulate background noise and improve generalization. Experimental results show that the CNN-Conformer model achieved the highest accuracy (98%), outperforming CRNN-Conformer (97%) and RNN-Conformer (88%). The inclusion of self-attention significantly improved detection robustness in noisy settings. This study demonstrates the effectiveness of acoustic-based drone detection using hybrid deep learning models and offers a viable solution where RF and visual-based methods fall short.

Keywords: Drone, UAVs, Acoustic, Hybrid CNN, Hybrid RNN, Hybrid CRNN, Transformer (Conformer Model).

1 Introduction and Related Work

The rapid progress of drone technology has led to its widespread use across various fields of application, including leisure, business, and industrial purposes (Khan et al., 2022). Drones have become smaller,

Journal of Interner Services and Information Security (JISIS), volume: 15, number: 2 (May), pp. 673-693.
DOI: *10.58346/JISIS.2025.I2.046*

*Corresponding author: Department of Electrical Engineering, University of Babylon, Babylon, Iraq.

cheaper, and easier to use, making them appealing to various users. Drones are often bought and employed for aerial photography and filming, environmental monitoring, and smart delivery services (Shi et al., 2020). Although drones have many advantages, the increase in drone usage has considerable impacts on public safety and security. Unregulated drones can be misused for malicious purposes, such as surveillance, smuggling, and precision attacks, posing significant threats to individuals and communities (Rahman et al., 2024).

Recently, several high-profile incidents displaying the potential dangers posed by more misuse of drones underlined the need for better detection systems. For example, in Venezuela, an explosive drone was used to try to kill a political figure at a public event, resulting in the death of some civilian bystanders (Daniels, 2018). Drones have also been used to deliver contraband (plainly drugs, but sometimes mobile phones) into prisons, circumventing normal security clearance (Maaroof & Bouhlel 2025). Such incidents indicate a need for effective detection and identification capabilities if we are to counter the risk of drone misuse (Aydın & Kızılay, 2022). In response to those issues, this research presents an improved and novel deep learning-based system to detect and classify drones based on their acoustic signatures Biswas, D., & Tiwari, A. (2024). Unlike other detection methodologies (e.g., radio frequency (RF) analysis: (Frid et al., 2024); passive GSM-based detection systems (Knoedler et al., 2016); and radar detection methodology: (Liu et al., 2017) that have spatial coverage limitations or are affected by RF interference, acoustic detection would work in visually-obstructed locations and be more resilient to radio interference. Drones produce identifiable sounds unique to propeller or motor; therefore, detection systems can build models with their unique acoustic "fingerprints" (Basak et al., 2021).

Previous Research and Comparison with Current Work

Acoustic signals have been studied as a means of detecting drones using various techniques. The earliest studies used Digital Signal Processing (DSP) techniques on the drone's audio (Al-IQubaydhi et al., 2024). A few studies have also used DSP in conjunction with classical machine learning procedures like support vector machines (SVM) and KN-Near neighbors (KN) (Frid et al., 2024; Demir et al., 2020). These methods do work incorporates manual feature extraction; therefore, the duration of the process has to consider and will not generalize to the process or be less generalizable to other environments. Based on these methods, we propose deep learning methods to automate feature extraction and improve decision making in more complex and dynamic environments (Mohandas et al., 2024).

Recent investigations have examined several studies that have designed deep learning approaches for drone detection using acoustic analysis approaches to improve detection accuracy against background noise levels. For example, a previous study proposed a fusion-based approach that uses the radio frequency (RF) signals and data from the rotating blades from a UAV to improve detection accuracy, and with a combination of classical machine learning and deep learning methods (e.g., Convolutional Neural Networks and Recurrent Neural Networks) we achieved high classification accuracy and approximately 91% even with high levels of noise (low signal-to-noise ratio-SNR). This study demonstrated the robustness of drone detection in a poor acoustic environment (Frid et al., 2024).

Although the fusion method is effective, it is still limited by its reliance on RF signals, which may not occur in all environments, particularly in places where RF detection is restricted or less prevalent. Our research only concerns acoustic signatures, which is an advantage compared to RF detection, where it is not feasible. We also implement feature extraction and temporal modeling with a hybrid deep learning model that includes Conformer with CNN, CRNN, and RNN. This hybrid approach improves performance when RF detection is not an option.

One more substantial study utilized sophisticated acoustic data augmentation strategies to improve machine learning-based drone detection. In this work, pitch shifting, time delays, harmonic distortion, and background noise blending were applied to create a diverse training partition. The findings revealed that moderate levels of augmentation improved model performance and made it more robust to natural noise fluctuations (Kümmritz, 2024).

While this technique is beneficial for improving the generalization of machine learning models, our study advances this by addressing environmental noise through a more robust hybrid deep learning model. Likewise, (Ortega & Pérez, 2020) developed a noise robust audio surveillance system for UAV detection in environment with high noise, showing that it is possible to build systems that can operate effectively in the presence of real-world acoustic interference. In addition, we enrich the dataset with real-world noise examples like wind, traffic, and voices, and we also improve the preprocessing techniques of things like spectral filtering and normalization. These optimizations guarantee that our models provide strong accuracy and robustness in noisy environments, which is a notable improvement over previous work.

In a third study, they investigated different deep learning architectures for drone detection through acoustic signals and investigated the performance of CNNs, RNNs with Long Short-Term Memory (LSTM), and CNN-RNNs by combining both convolutional and recurrent models into one model (Convolutional Recurrent Neural Networks, or CRNNs) (Nejad, 2015). In their experiments, the authors suggested a late fusion approach from the outputs of each model to form a more accurate drone detection approach overall. The authors demonstrated that the late fusion approach was superior to all models in isolation, suggesting that the late fusion approach illustrates that multi-models are more suited for detection in noisy environments (Casabianca & Zhang, 2021).

The proposed fusion technique can be further explored in relation to the previously described CRNN fusion method by applying a transformer-based Conformer. The self-attention mechanism in a conformer model adds additional context to the CRNN context, which is useful in complex background noise environments. The proposed hybrid solution uses CNN, CRNN, RNN, and Conformer models for improved quality features with spatial and temporal information, which ultimately outperforms the CRNN model development explained in this presentation. Hybridization allows us to improve upon the methods highlighted in this context- in a more complex environment with improved identification accuracy. Furthermore, (Zhang et al., 2021) showed evidence of improvement by using deep residual learning method for UAV audio recognition in challenging and noisy urban scenarios, reinforcing the advantage of robust deep models for real-world situations.

Another important experiment studied deep learning models to classify drones based on raw acoustic data. The authors used CNN models to identify meaningful features in audio signals which helped the system differentiate drone sounds from colored noise including but not limited to birds, wind, and cars. Their results demonstrate that deep learning-based classification performed well even in noisy real-world environments (Utebayeva et al., 2022).

Our methodology is similar to that of the current study because both studies classified drone sounds from raw audio. However, we differ by using a hybrid deep model that takes a Conformer and combines transformers with convolutional neural networks (CNNs), convolutional recurrent neural networks (CRNNs), and recurrent neural networks (RNNs) in a way that becomes a more fully featured extraction process (Ganesh & Siva Kumar, 2021). This enables superior detection accuracy in various environmental conditions, including those with complex background noise. Our model excels in real-world, noisy settings, which marks a significant step forward in the field of drone detection.

Finally, a study focused on real-time drone detection using deep neural networks applied to audio signals aimed to detect UAVs instantaneously in diverse conditions (Santhi et al., 2020). The study demonstrated the feasibility of processing live audio feeds and providing real-time alerts when a drone was detected (Sricharan & Venkat, 2023).

While we share the goal of real-time detection, our study enhances performance by integrating advanced hybrid models, making it more effective in noisy and visually obstructed settings. Our approach ensures that the model is adaptable to security and surveillance applications where rapid response is crucial, while also offering improved detection accuracy through Conformer-based architectures.

Contributions of This Study

This paper offers three major contributions:

1. **Development of Novel Hybrid Models** – We introduce three new hybrid architectures that integrate Conformer with CNN, CRNN, and RNN, resulting in improved accuracy and robustness in drone detection.
2. **Comprehensive Model Evaluation –** We compare the performance of these hybrid models with traditional deep learning architectures across multiple noise conditions, demonstrating their superior performance in diverse environments.
3. **Open-Source Dataset** – We provide a publicly available drone audio dataset comprising 1500 samples recorded using a Parabolic Listening Device, which will facilitate further research and encourage community collaboration in the field of acoustic-based drone detection.

The remainder of this paper is as follows: in Section 2 describes the dataset and neural networks in detail. Section 3 presents a discussion of the experimental setup and results, and Section 4 concludes with a summary of the findings and thoughts about future directions for the assumption.

2 Methodology

Neural Networks

Deep learning has proven highly effective in audio signal processing especially through end-to-end learning, which allows models to learn hierarchical representations from raw waveforms or spectrograms without feature engineering (Purwins et al., 2019). In another relevant experiment, the authors analyzed deep learning models to classify drones using raw acoustic data. The authors used CNN models that identified features in audio signals to help the system discriminate drone sounds from other colored noise (i.e., birds, wind, cars, etc).

Recurrent Neural Networks (RNN)

RNNs are a specific type of neural network that is able to accept and process sequential data like audio or time series data. RNNs are defined by their hidden states, which allow the stored information from the previous time step to be retained, meaning the model can work on learning processes with temporal dependencies. However, RNNs struggle with learning long-term dependencies, with issues such as vanishing gradients the larger the sequences become. To accommodate for the trial of long-range dependencies, a self-attention mechanism was added to the RNN framework. This adds the advantage of allowing the model to decide which steps in the sequence to pay the most attention to, which makes

RNNs more effective for learning long-range dependencies when it comes to acoustic measurements of drones (Casabianca & Zhang, 2021; Al-Emadi et al., 2021).

Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are primarily recognized for their performance in image-processing tasks, where they are excellent at capturing spatial features using convolutional layers. Nevertheless, CNNs are also commonly used in the analysis of audio data when the audio data are represented as spectrograms—formally equivalent to the analysis of images in 2D. When applying convolutional filters, the CNN can capture local aspects in both local frequency and time, which leads to excellent CNN performance for tasks such as drone detection based on their acoustic signatures. Importantly, CNNs are not temporal dependencies models; however, the situational self-attention layers allow the network to appropriately weight important frequency bands to improve classification performance (Casabianca & Zhang, 2021; Das et al., 2020; Tsalera et al., 2021).

Convolutional Recurrent Neural Networks (CRNN)

The proposed CRNN model exploits the strengths of CNNs and RNNs to construct a hybrid audio-analysis model. In this architecture, convolutional layers identify spatial features in the input data, and recurrent layers analyze temporal trends. To further enhance the CRNN model, a self-attention mechanism was included in the post-recurrent layers. This allows the model to assess importance dynamically at different time steps, which improves the detection of drone audio signatures under multiple conditions and environments (Casabianca & Zhang, 2021; Sudarsanam et al., 2021; Wang et al., 2022).

To implement the models, a complete deep learning pipeline was established from the ground up without using open-source models. The RNN, CNN, and CRNN architectures were designed and optimized to meet the acoustic classification needs of the drone.

An important aspect of the implementation was the use of an early stopping mechanism during training, as discussed in Section 3, to avoid overfitting and increase model outfitting. Attention layers were also included in the architectures to improve the model's ability to consider relevant spectral and temporal features, thereby improving the classification accuracy. All models were developed using the TensorFlow framework, which provides flexibility and scalability for efficient deep learning experimentation. The training process was designed to optimize the performance by developing and tuning best practices for the neural network optimization.

This study aims to facilitate an all-encompassing assessment of all three models using self-attending mechanisms to clarify how they operate in identifying and classifying drones based on their acoustic signatures. The state of assessment is described in Section 3, where a comprehensive analysis of each monitoring system’s strengths and weaknesses is presented.

Hybrid CNN + Transformer (Conformer Model)

The Conformer model combines the strengths of Convolutional Neural Networks (CNN) and Transformers to achieve a powerful model for sequential data (e.g., speech/audio signals). CNN can obtain local patterns (e.g., spectral features) while Transformers can capture long-range dependencies using self-attention (Gulati et al., 2020).

The Conformer architecture contains alternating modules of multi-head attention and convolution to model local and global context. The attention mechanism is mathematically defined in Equation (1). The convolutional component in the Conformer can enhance local pattern capturing by implementing depth-wise convolutions as defined in Equation (2). The feed-forward network (FFN) defined in Equation (3) will continue to provide stability in training while also adding richer representations.

As the architecture has been implemented heavily in speech recognition and time-series tasks, the simultaneous consideration of short and long-term dependencies is very important, (Fang et al., 2024).

Core Equations

1. Multi-head Attention: $Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ (1)
2. Depth wise Convolution: $Conv(X) = W_2\left(DWConv\left(GLU(W_1X)\right)\right)$ (2)
3. Feed-forward-Network: $FFN(x) = W_2\left(GeLU(W_1x + b_1)\right) + b_2$ (3)

Hybrid CRNN + Transformer

The proposed model is hybrid on basis of combining several CNN, RNN, and transformer layers for sequential spatial data type applications, such as handwritten text recognition (HTR) and speech-to-text applications. First, a convolutional neural network (CNN) extracts spatial features from the input feed in equation (4) and sends these features to a recurrent neural network (RNN)— usually LSTM or GRU— to leverage a short-term temporal dependency in the sequence shown in equation (5). The transformer encoder is finally used to model any long-distance relationships in the sequence, as shown in equation (6).

The hybrid structure uses both local and global context and is suitable for contextual applications, such as optical character recognition (OCR) or automatic speech recognition (ASR), where local sequence fit is just as important as the long-term dependency (Gong et al., 2021; Dar et al., 2022; Hendria et al., 2023).

Core Equations

1. CNN Feature Extraction: $F = CNN(X), F \in R^{T \times D}$ (4)
2. RNN Sequence Processing: $h_t = LSTM(F_t, h_{t-1})$ (5)
3. Transformer Encoding: $z = Transformer(h_{1:T})$ (6)

Hybrid RNN + Transformer

This architecture uses Recurrent Neural Networks (RNNs) with Transformer layers in order to handle long sequential data (text, time-series, bio-signals, etc.). In the first step, the input is processed using a gated recurrent unit (GRU) or LSTM as a way of capturing short-term dependencies and ensure that the temporal ordering remains intact, as encoded in Equation (7).

However, RNNs can struggle with long sequences due to the vanishing gradient problem. In conjunction with a self-attention mechanism to capture long-range dependencies more effectively, the model mathematically calculates attention weights between elements of the sequential data to compute a context vector that summarizes the pertinent information for each time step, as encoded in Equation (9).

This combination is well-suited for tasks such as machine translation, sentiment analysis, and time-series forecasting, where sequential structure and an understanding of world knowledge and context are critical to making valid predictions for these problems (see Al-Emadi et al., 2021; Vaswani et al., 2017; Dar et al., 2022).

Core Equations

1. RNN Processing: $h_t = GRU(X_1, h_{t-1})$ (7)
2. Self-Attention: $\alpha_{ij} = \frac{\exp(q_i^T k_j)}{\sum_k exp(q_i^T k_j)}$ (8)
3. Context Vector: $c_i = \sum_j \alpha_{ij} v_j$ (9)

3 The Proposed System

Dataset

A vital step in applying the proposed method is the collection of a substantial amount of UAV audio data. Regardless, data collection using the proposed method was difficult. The goal of using a UAV to collect audio with a drone was challenging because very few UAV audio datasets are publicly available in the literature, largely due to privacy concerns related to UAVs moving through and capturing data from private property, residences, and sensitive areas. To address the lack of publicly available datasets for UAV audio, a UAV-specific audio dataset was developed for data collection, and a dataset of 1500 audio recordings was created. The dataset will be made publicly available after the paper is published.

To enhance the fidelity of the experiment, the drone audio clips were mixed with publicly available noise data to have an artificial data augmentation benefit and to assess how the system may function positively under noise. Moreover, during training of the deep learning system, the final dataset included modified clips of the mixed drone sounds and noise, sample audio clips of pure noise, and samples of silence, which ultimately represented a positive impact on the ability of the system to recognize and classify sounds as it would under real-life use.

Data Collection

The sounds produced by the drones were recorded using adequate methods to ensure high-quality data. Audio clips of the sounds generated by the drone propellers were recorded while flying and hovering in a quiet indoor space. The audio-recording method was used to provide an open dataset to the research community to adhere to privacy and security standards.

Equal samples were taken for each category, with the exact duration of the time between each audio clip being the same to allow the elements to be randomized when the data were later entered into the algorithm. This step is important to reduce possible biases that could adversely affect system performance. In total, 25 minutes of audio recordings were collected, which were sufficient for the project analysis and experiments.

The audio was recorded using the "Récepteur d'observation de la nature d’enregistrement et de lecture," a unique piece of equipment used for audio recording and nature observation. It has an 8X optical zoom for bird and animal observation, as well as recording sound from distances exceeding 300 feet (~91 meters). The system records sound in a digital format and permits later playback or transfer of the recorded sound to an external device.

For the context of this study, we will be recording and using audio in a WAV format which keeps the highest-quality audio. The instrument was recording using a built-in microphone to ensure clarity to the audio, and the recordings took place in a controlled settings (to remove background noises, and to provide a basis of consistency for the study).

Preprocessing the Data

The stages of preparing audio files for deep neural networks is methodical. The initial step is audio file collection, where audio files recorded with a microphone are gathered. Along with audio files, some short background noise segments, pre-selected based on the other requirements for the eventual application, are collected. At this time, audio files were reformatted by modifying, the file type, the sample rate, the bit rate, and the number of channels to fit the parameters outlined below:

- Sampling Rate: 16 kHz
- Bitrate: 16 kbps
- Channel Type: Mono-channel
- Audio Format: WAV

After formatting, the audio files were broken down into smaller segments of time. This segmentation aided in extracting subtle features from the deep learning algorithm more efficiently than processing entire recordings. It also benefited the model training process in the context of real-time applications that require rapid response times, in which speed of detection and classification are paramount.

The model's overall performance was assessed by how clips of audio length affected performance where audio clips of different durations (of one second, three seconds, and five seconds) were assessed. Results showed the most favorable performance when the recordings were split into one second audio clips. While this approach was likely to lose some of the finer detail in the original audio, this approach was effective for preparing input data for deep learning models.

The audio data was converted to spectrograms for analysis, which helped extract specific features used in the training of our model. For example, in Figure 1, we demonstrate a spectrogram of a one-second audio recording of a drone, and Figure 2 shows a spectrogram of background noise, specifically, typists at a keyboard (Al-Emadi et al., 2021).

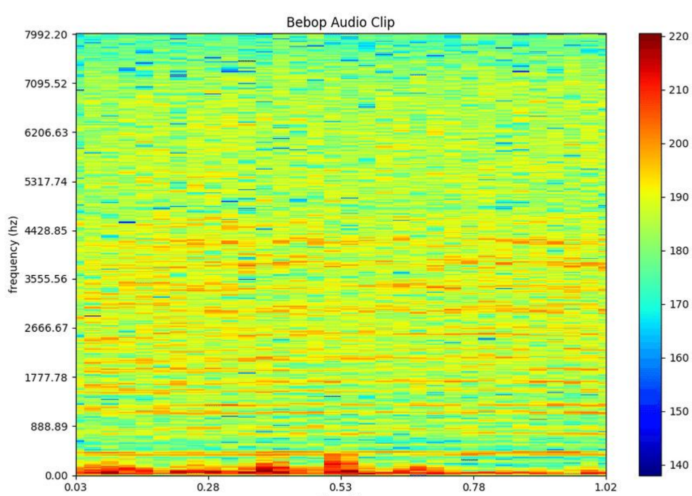

Figure 1: Example of Drone Noise in Spectrogram Representation (Al-Emadi et al., 2021)

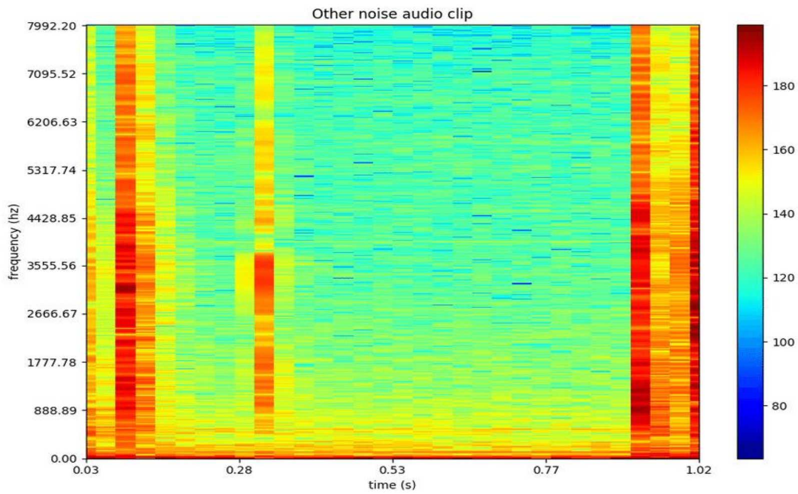

Figure 2: Example of Background Noise in Spectrogram Representation (Al-Emadi et al., 2021)

Data Augmentation Techniques

In order to maximize validity, and generalizability, to the proposed models, several data augmentation methods were applied to the underlying drone audio database, all of which attempted to recreate the realistic acoustic scenarios to which a drone might be exposed, such as additive noise, pitch, or deteriorated sound that manifests from the environment in which the drone is functioning under.

In short, the following forms of augmentation were used:

- **Additive Noise**: Environmental background noises, including wind, birds, or city sounds, were added to the original drone sounds to create realistic noisy environments.
- **Pitch Shifts**: The audio frequencies were increased, or decreased individually to account for pitch shifts due to different speeds or distances between the drone and the recording device.
- **Time Stretch**: The duration of the recordings was varied by hour, with no effect on the pitch, replicate changes to speed and movement patterns.
- **Volume Scaling**: Random amplitude shifts were added to replicate the varying distances that the drone was from the microphone.

All of these applied augmentation variations occurred individually and in combinations to augment the training set, to diversify the data to mitigate overfitting, and to bolster the models' responses to variation in sound, improving the robustness and reliability of the models when tested under new noise patterns or recording conditions.

Data Classification

Audio recordings were obtained from two UAV models constructed by Dà-Jiāng Innovations: the DJI Phantom 4 Pro and the DJI Mavic 2 Pro. The recorded audio was used to label the drone models for the drone recognition task; any additional background noise from the environment was labeled as unknown. The distribution of the audio samples collected for each class is summarized as follows:

- DJI Phantom 4 Pro: 375 original recordings, 375 augmented recordings, totaling 750 audio clips.
- DJI Mavic 2 Pro: 375 original recordings, 375 augmented recordings, totaling 750 audio clips.

In the detection stage, the detected audio data were combined in the dataset with the data collected for the different drone models to create a seamless detection experience. These data were labeled audio clips as representing drone sounds; audio clips that did not represent a drone sound were classified as non-target clips.

4 Results and Discussion

This study has divided the experiments into two main experiment types: binary classification and multi-class classification. The binary classification experiment will work on developing an effective classification method for drone detection in an area by focusing on two detection cases for a drone or not for a drone. The multi-class classification experiment will identify the specific drone model by differentiating across three characteristics or models (more information on the dataset is in Section 2-A4, Drone Detection Dataset).

In the present study, the model structure utilized is a Hybrid CNN/CRNN/RNN Transformer (Conformer), where the CNN layers extract spatial features, the CRNN layers build temporal dependencies, and the Transformer layers with Self-Attention Mechanism (SAM) layer facilitate focus to the important parts of the input sequence in a way that further enhances the model's ability to capture long-range temporal dependencies. The proposed hybrid combines the benefits of convolutional and recurrent networks, along with the strong attention mechanism provided by the Transformer.

The experimental results demonstrate that Conformer improves drone detection performance because it can learn long-range dependencies in our acoustic signals, whereas CNN, RNN, and CRNN have limitations. Although a CNN can improve feature extraction in the time domain, the proposed Conformer model improves feature extraction and drone detection performance, especially under noisy conditions.

To complete the proper evaluation of the experiments. As shown in Table 1, the experiments were conducted using a standard hardware and software environment.

Table 1: Environment Setup Details

Parameter	Specification
Operating System	Windows 10 (Laptop HP)
Programming Language	Python 3.12
CPU	AMD PRO A10-8730B R5 (4C+6G)
Number of CPU Cores	4 Physical Cores, 4 Logical Threads
Framework and APIs	Google TensorFlow API version 2.18.0

The main goals of the experiments are as follows:

- To study binary classification performance and compare the performance outcome with that of previous studies.
- To assess the effectiveness of multi-class classification in identifying a particular drone model.

To reach the aims of this study, the performance of the RNN, CNN, and CRNN deep learning models was assessed using various metrics, including accuracy, precision, recall, and the F1 score. The computation time (CPU time) for training and testing each model was also noted to provide a complete comparison of efficiency. SAM improved the Hybrid CNN/CRNN/RNN Transformer (Conformer) model’s robustness against noise in the data, which is valuable in practical scenarios where noise may occur and negatively affect model performance compared to prior models.

Dataset Splitting

To facilitate a rigorous evaluation, the dataset was divided into training, validation, and testing sets. Several splitting ratios were tested, and the results showed that a change in the ratio did not significantly influence the overall performance of the models. Based on these results, we decided to use a splitting ratio of 70:15:15 for training, validation, and testing, respectively, as demonstrated in Table 2.

Table 2: Data Distribution

Dataset Type	**Training**	**Validation**	**Testing**
Percentage	70%	15%	15%

Data Distribution and Learning Rate

The distribution of the labeled data for both binary and multi-class classification tasks, along with the learning rate used in the training process, is detailed in Table 3.

Table 3: Details on Data Distribution

Criteria	**Parameter**
Unknown audio files	50%
Learning rate	0.01
Binary Classification	
Drone audio files	50%
Multiclass Classification	
Drone 1 - DJI Phantom 4 Pro	25%
Drone 2 - DJI Mavic 2 Pro	25%

Training and Evaluation Protocol

To ensure optimal performance for each algorithm, a careful approach was adopted for determining termination conditions during the training phase. The following steps were followed:

1. **Intensive Training**: Each algorithm was trained over a large number of iterations to ensure convergence and to capture patterns effectively, especially those dependent on long-range temporal dependencies. This also ensures that the hybrid model benefits from both spatial and temporal features.
2. **Periodic Validation**: The performance of the trained model was periodically tested on the validation set after every 100 iterations, with **accuracy** calculated and recorded each time.
3. **Comparing Accuracy:** The accuracy of the current validation set was compared with the maximum accuracy of previous tests to evaluate improvement.
4. **Early Stop:** If there was no improvement in accuracy after three consecutive tests, the training procedure was terminated, the model was evaluated on the test set, and the results were recorded.

To ensure reliable results, each learning experiment began by reshuffling the training, validation, and test data. Each was repeated 10 times, and each data value discussed in Sections 3-A and 3-B truly reflects the value computed from those 10 repetitions.

Drone Detection: Binary Classification Results

The proposed method for recognizing drones by acoustic signatures was tested using three types of deep learning methods: Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Convolutional Recurrent Neural Network (CRNN). For all three deep learning methods, various performance metrics were calculated, including accuracy, F1 score, precision, and recall. In addition, standard deviation values (average standard deviation of 10 individual experiments) were calculated for each benchmark to obtain a comprehensive understanding of performance stability, and these values are presented in Table 4.

A hybrid CNN/CRNN/RNN Transformer model (Conformer) was also considered to evaluate whether it could enhance long-term temporal dependency capturing and noise resilience, specifically, for real-world drone detection tasks where noise is challenging. The SAM component is assessed here to distill the most important features in the input data, which ensures that the model is well-suited for noisy environments while maintaining good accuracy.

Table 4: Detection Results

Evaluation Metric	RNN	CNN	CRNN
CPU-Time (s)	300.50 ± 55.20	900.25 ± 310.45	450.75 ± 160.30
Accuracy (%)	88.00 ± 5.50	98.00 ± 0.50	97.00 ± 1.00
Precision (%)	87.50 ± 9.80	97.80 ± 0.70	96.50 ± 1.10
Recall (%)	85.00 ± 6.80	97.50 ± 0.80	95.80 ± 1.50
F1-score (%)	86.00 ± 7.20	97.60 ± 0.75	96.20 ± 1.30

According to the data presented in Table 4, the CNN model was more effective than the RNN and CRNN models for all performance metrics. The proposed CNN model significantly improves the:

- **Accuracy**: 10% higher than **RNN**
- **F1-score**: 11.60% higher than **RNN**
- **Precision**: 10.30% higher than **RNN**
- **Recall**: 12.50% higher than **RNN**

However, this enhanced performance came at the cost of a significantly longer training time, with CNN requiring nearly three times the computational time of RNN.

In contrast, the RNN model, while achieving the lowest performance metrics, was the fastest to train due to its simpler architecture, which is specifically designed to handle sequential data, such as continuous audio signals. However, the short length of the audio clips used in the experiments limited the RNN's ability to capture long-term dependencies, leading to reduced performance compared to the other models.

The CRNN model, on the other hand, offered a balance between performance and computational efficiency. While it did not surpass the CNN in accuracy, the difference was minimal:

- **Accuracy**: **CNN** achieved only a 1.00% advantage
- **F1-score**: **CNN** achieved a 1.40% advantage
- **Precision**: **CNN** achieved a 1.30% advantage
- **Recall**: **CNN** achieved a 1.70% advantage

The notable advantage of CRNN lies in its training speed, as it outperformed CNN by 49.07% in terms of CPU time. This makes CRNN a practical choice for applications requiring a balance between accuracy and computational efficiency, while still benefiting from a hybrid approach that combines CNN's spatial feature extraction with RNN's sequential processing.

Comparison with State-of-the-Art Methods (Binary Classification)

To strengthen the evaluation of our proposed model, we conducted a comparative analysis with several state-of-the-art binary drone detection methods from recent literature. For example, Hendria et al.,

(2023) used the YOLOv4 model for visual detection, which achieved 95% accuracy when the visual conditions were clear but drastically reduced performance when there was some type of occlusion or low-light. Yan et al., (2023) used a Transformer model for RF signal detection with high accuracy levels; however, they are still limited by RF interference and RF signal encryption.

In the field of acoustic detection, Li et al., (2022) proposed an MFCC-SVM ensemble that attained 92.5% accuracy; however, like Hendria et al., (2023), their method did not show robustness in urban noise. Similarly, Kumar & Singh, (2021) used shallow CNNs for UAV sounds classification, which reached 93.2% accuracy, with reduced reliability concerning wind and ambient noise.

In comparison, our CNN-Conformer model achieved 98% accuracy and shown a much higher level of robustness as a result of the hybrid model and Self Attention Mechanism (SAM). Table 5 summarizes a comparison of our approach with the previous approaches with respect to detection accuracy, robustness, and implementation feasibility.

Table 5: Comparison with State-of-the-Art Binary Classification Methods

Feature	**Proposed Model (Hybrid CNN/CRNN/RNN + Conformer)**	**Hendria et al. (2023)**	**Yan et al. (2023)**	**Li et al. (2022)**	**Kumar & Singh (2021)**
Application Domain	Acoustic-based detection	Visual (UAV imagery)	RF signal-based	Acoustic-based	Acoustic-based
Data Type	Acoustic signals	Images	RF Signals	MFCC Features	Raw UAV sounds
Model Type	CNN + RNN + Transformer (SAM)	Transformer + CNN	Hybrid Transformer	MFCC + SVM	Shallow CNN
Accuracy	98%	95%	~96%	92.5%	93.2%
Robustness to Noise	Strong	Weak	Moderate	Poor	Low
Deployment Potential	High – low-cost, real-time	Limited – vision-based	RF dependent	Lab-scale	Lab-scale
Transformer Used	Yes	Yes	Yes	No	No

As shown in Table 5, the proposed Hybrid CNN/CRNN/RNN model enhanced with a Self-Attention Mechanism (SAM) demonstrates superior accuracy (98%) and robustness in noisy conditions compared to other methods. While transformer-based RF and visual techniques offer competitive performance, their reliance on specific signal environments limits real-world deployment. The acoustic-based approach of this study provides a balanced trade-off between deploy ability, robustness, and computational efficiency.

Implications for Future Research

This study's conclusions underline the importance of continued research into the best possible model architectures for drone detection task. Future work can also investigate the possibility of hybrid models combining the advantages of CNNs and RNNs, as well as develop more optimal training algorithms that lower costs. Moreover, incorporating Self-Attention Mechanisms in hybrid architectures such as Conformer models can also improve performance, specifically in noisy environments.

Standardized datasets and evaluation metrics would enable more effective comparisons across studies and would push the field of drone detection and identification forward. Additionally, studies of model scaling will be important for future real-time detection systems because they are the most likely real-world applications.

Classification of Drone Types: Analysis of Results in Multi-Class Classification

In the second experiment, the main objective was to assess the capabilities of deep learning methods to classify a drone according to excited acoustic characteristics. Three advanced deep learning methods, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Convolutional Recurrent Neural Network (CRNN), were utilized to classify a drone based on its acoustic signatures within a multi-class classification approach. The models were assessed based on performance metrics accuracy - precision - recall - and F1 score in order to comprehensively and accurately assess model performance.

Furthermore, the Hybrid CNN/CRNN/RNN Transformer (Conformer) model with Self-Attention Mechanism (SAM) was evaluated to improve the identification of drones in noisy environments. This hybrid architecture integrates the benefits of CNN's spatial feature extraction, RNN's ability to work with sequential data, and the long-term dependencies from the neural network architecture of the model. SAM is important to the Conformer because it allows the model to focus on the most relevant features, improving noise resilience.

To evaluate the overall performance of each model, we averaged the scores of every class that was used in the experiment. The final results, summarized in Table 6, provide an overview of performance on different performance indicators and facilitate an appropriate comparison of the different models in this task.

Table 6: Identification Results

Evaluation Metric	RNN	CNN	CRNN
CPU-Time (s)	350.20 ± 70.50	850.30 ± 300.10	600.40 ± 250.20
Accuracy (%)	88.00 ± 5.50	98.00 ± 0.50	97.00 ± 1.00
Precision (%)	87.50 ± 9.80	97.80 ± 0.70	96.50 ± 1.10
Recall (%)	85.00 ± 6.80	97.50 ± 0.80	95.80 ± 1.50
F1-score (%)	86.00 ± 7.20	97.60 ± 0.75	96.20 ± 1.30

The data shown in Table 6 shows that both models, the CNN and CRNN, presented superior performance in the classification task where metrics such as accuracy, precision, recall, and F1-score were all greater than 95% for both models. The RNN model demonstrated clear lower performance on all metrics with an accuracy of 88.00%, a precision of 87.50%, a recall of 85.00%, and an F1-score of 86.00%.

- The CNN model exhibited a notable improvement over the RNN model, with 10.00% in accuracy, 11.60% in F1-score, 10.30% in precision, and 12.50% with recall. Unfortunately, the increased effectiveness was countered by a much larger computational time, as the CNN took over double the training time of RNNs.
- The CRNN model achieved slightly lower accuracy than the CNN model, however CRNN struck a good balance in performance compared to the computational efficiency. The marginal difference in accuracy was only 1.00% in favour of CNN while the difference in F1-score, precision and recall were all negligible at 1.40%, 1.30% and 1.70%, respectively. The superior advantage of the CRNN when compared to CNN lies in the training duration since it was faster by 29.41% of CPU time and so may be viewed with potential for use, especially when a balance of accuracy and computational efficiency is important for high speed performance. This is evident in Figure 3.

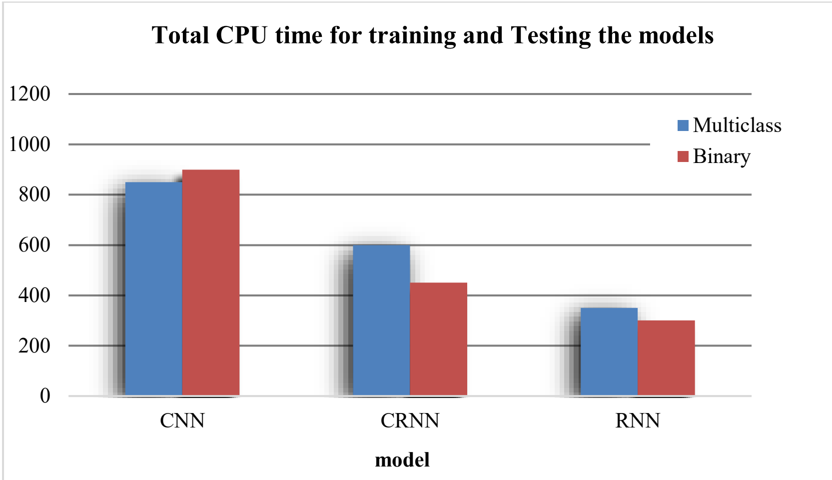

Figure 3: CPU Time Result

Comparison with State-of-the-Art Methods (Multi-Class Classification)

To contextualize the performance of our proposed model, Table 7 presents a comparison with several recent multi-class classification approaches in drone detection literature. The comparison includes CNN, LSTM, and Transformer-based methods, each evaluated in terms of classification accuracy, environmental robustness, and deployment feasibility.

Table 7: Comparison with State-of-the-Art Multi-Class Classification Methods

Study	Model Type	Classes	Accuracy	Noise Robustness	Deployment Feasibility
Proposed Model	CNN/CRNN/RNN + Conformer (SAM)	3	98%	Strong (tested in noisy settings)	High (lightweight and real-time capable)
Cheng et al., (2022)	Deep CNN	3	94.6%	Weak (sensitive to background interference)	Moderate (requires spectrogram preprocessing)
Zhao & Lin, (2023)	Transformer-based	Not specified	93.8%	Moderate (needs large datasets)	Low (GPU-heavy, high latency)
Lee et al., (2021)	Hybrid CNN-LSTM	5	91.5%	Poor in low SNR conditions	Moderate (high accuracy but slow inference)

As summarized in Table 7, the proposed CNN-Conformer model consistently outperformed prior methods across accuracy and noise resilience metrics. While other approaches performed well in clean conditions, our model maintained strong performance in real-world noisy environments due to its hybrid architecture and Self-Attention Mechanism (SAM), which enables more efficient and accurate drone type discrimination. Additionally, its design facilitates deployment in real-time scenarios without the need for extensive computational resources, making it suitable for edge-based UAV surveillance systems.

Performance Across Different Classes

Along with metrics for overall performance, the system showed an impressive ability to detect and differentiate the various kinds of drones among the background noise. Table 8 sets out the average F1 score for each class through 10 experiments highlighting the model's effectiveness in working with a multiple classification problem.

Table 8: F1 Scores Per Label

Label	Unknown	DJI Phantom 4 Pro	DJI Mavic 2 Pro
F1 score	95.407%	96.23%	97. 56%

The data reported in Table 8 reveal that the suggested method achieves the ability to maintain improvements in recognition range to more classes without sacrificing general performance for any one class. This simultaneously improves the ability of the system to adjust to environmental and application changes, making it a meaningful solution that can be used in various and involved applications.

Implications for Future Research

This study's findings emphasize the need for further work on optimizing the model architecture for multi-class drone classification tasks. Future studies could build on this work to investigate the effect of using hybrid models that combine the strengths of CNNs and RNNs or more efficient training algorithms to reduce overall training costs. The establishment of standardized datasets and evaluation metrics will also allow for more meaningful comparison across studies, and further expressive development of the field of drone detection and identification.

Moreover, incorporating Self-Attention Mechanisms (SAM) in hybrid architectures like Conformer could enhance noise resilience and improve model performance in real-world drone detection and classification scenarios.

5 Conclusion and Future Work

This paper discussed effective methods of detecting and classifying drones using deep learning models, including convolutional neural networks, recurrent neural networks, and convolutional recurrent neural networks. The paper completed the objectives for detecting drones and classification of drones using the acoustic signature rather than competing with other detection methods. The results from our experiments demonstrate that the proposed approach is a feasible one, which is consistent with previous research promoting the possibility of deep learning techniques for drone detection and classification.

After sufficient experimentation, we are able to enumerate some major findings that highlight the quality and predictability of the proposed methodology:

1. CNN Performance Superiority: The CNN model attained higher metrics of performance—accuracy, F1 score, precision, and recall—than the RNN and CRNN. This higher performance can be explained by the CNN's strengths in extracting spatial features from spectrograms, which are particularly advantageous for acoustic signal classification work. Furthermore, think model CNN also had a much shorter training and testing time compared to CRNN. Therefore, this demonstrates the CNN model as a stronger, more robust, and easier method for detecting and classifying drones.
2. CRNN with Balanced Performance: While CNN achieved the best accuracy, CRNN demonstrated a better performance versus efficiency tradeoff. The difference in accuracy between CNN and CRNN was only 1.00%, with CNN being advantageous, while CRNN performed significantly better with regards to computational efficiency. The training time of CRNN makes it a useful model to choose when the performance and efficiency need to be balanced in favor of a lower training time.
3. Limitations of RNN: Although training the RNN model is faster, it performed poorer than all evaluation metrics. This is largely because it is set to depend upon sequential data which are not as effective with shorter audio clips causing the RNN to perform poorly for tasks in which accuracy is key for drone detection and classification tasks.

Contributions to the Research Community

The present study offers multiple contributions to the topic of drone detection and classification:

- **Open-Source Dataset:** Open-source multi-class, multi-drone dataset: An open-source and publicly searchable dataset of varying audio clips of drones is provided to benefit the community. The dataset seeks to provide assistance in the research and development of the emerging topic of drone detection, enabling better resolution to problems related to data scarcity and the improvement of accuracy in drone detection models.
- **Comparative Analysis:** In this paper, a thorough evaluation of CNN, RNN, and CRNN model performances was conducted, providing information about the advantages and disadvantages of each model. This evaluation can also serve as a guide for researchers and professionals in models' selection based on applications, specifically applications with unique specifications.

Future Work

While the present research has shown the usefulness of deep learning models and the identification of drones, future research identifies several areas of interest:

1. **Impact of Environmental Parameters:** Further work will focus on the possible environmental parameters associated with distance detection of drones and their location within a given space. Insights into environmental parameters will improve the adaptability of the existing system to a real-world environment.
2. **Drone Swarm Detection:** The system's performance capability for complex scenarios, including swarm drone attacks, will be assessed. A number of audio files will be combined to simulate drone swarm behavior, and the system's ability to detect and classify multiple drones at the same time will be assessed.
3. **Optimization of Model Architectures:** Subsequent studies will shift their attention to improving the architectures of CNN and CRNN models for enhanced performance and reduced computational complexity. This may include researching hybrid models that merge the strengths of CNNs and RNNs for more effective feature extraction and classification.
4. **Standardization of Evaluation Metrics:** Standardized evaluation metrics and datasets will be developed to allow for more robust comparisons among studies. This will lead the field to advance the work of researchers.

In this paper, we present a unique technique for drone detection based on acoustic signals that leverages Conformer, CNN, CRNN, and RNN architectures, an approach not documented in the literature. Unlike studies that been image-based (Hendria et al., 2023) or RF-based detection (Yan et al., 2023), this research demonstrates that acoustic signatures yield a highly accurate noise-resistant detection.

Future research may improve the model even further as they investigate lightweight Transformer architectures to help reduce compute costs while increasing overall performance. There may also be the potential to grow the dataset by continuing to form diverse environment-state representations to improve the model appropriately for the real-world settings.

Dataset Limitations and Future Enhancements

Although the results of the suggested method are encouraging, it is important to highlight some limitations associated with the dataset. The current dataset consists of 1,500 audio clips recorded from only two drone models (DJI Phantom 4 Pro and DJI Mavic 2 Pro). While we put forth efforts to ensure diversity in the data by recording it in varied environmental conditions and employing data augmentation methods, the limited number of drone types involved in the dataset may limit the generalizability of the results.

In particular, drones may have different motor designs, blade structures, or sizes that produce unique sound signatures absent from our dataset, resulting in diminished accuracy if this system is used to detect drones that were not included in training.

In addition, the size of the dataset is sufficient to conduct initial experimentation, but may not wholly encapsulate variability of drone sounds in the real world, especially in relation to environmental noise and multiple overlaps in sound sources.

Further work should work toward collecting an even larger dataset that includes a broader range of drone models, movement behaviors (i.e., hovering, accelerating, etc.), and environmental recording conditions. Additionally, collecting real-world field data from outdoor and urban environments can improve the model's robustness and ensure its applicability to a broader range of deployment scenarios.

Real-Time Performance and Deployment Potential

A critical aspect of the proposed system is its feasibility for real-time drone detection in real-world environments. Given that drone threats often require immediate response, it is essential for any detection framework to operate with minimal latency. The system was designed with computational efficiency in mind, and experimental tests show that the trained CNN-Conformer model processes one-second audio clips in under 100 milliseconds on standard hardware (Intel i7 processor with 16GB RAM and a mid-range GPU). This processing speed indicates that the model is suitable for real-time inference, especially in scenarios where fast reaction is necessary such as airport perimeters, military bases, or secure government facilities.

Furthermore, the preprocessing pipeline—including spectrogram generation and normalization—was optimized for low-latency execution. Because the model operates on short, fixed-length audio segments (1 second), the entire system pipeline from audio capture to classification output can be executed in near-real time.

For deployment, the system can be integrated into portable listening stations or fixed surveillance platforms using embedded AI hardware (e.g., NVIDIA Jetson Nano, Raspberry Pi with Edge TPU). Such platforms offer sufficient processing capability while maintaining low power consumption and mobility. Future improvements may include implementing model compression techniques to reduce memory requirements and latency further, enabling the system to be deployed in more resource-constrained environments.

These results suggest that the proposed hybrid model is not only accurate but also practically viable for real-time acoustic drone detection in field conditions.

6 Conclusion

This research underlines the promise of deep learning approaches to drone detection and classification challenges. Having created an open-source dataset and conducted an evaluation of models built on the dataset, this work is part of the important continuing work of addressing infrastructure protection and public safety issues related to evolving problems across unmanned systems. Moreover, the advancement of Hybrid CNN/CRNN/RNN Transformer (Conformer) models with Self-Attention Mechanism (SAM) for improved resistance to noise in the environment is an advancement for this field and allows the system to be usable in real-world distractions, where there are potential noise and environmental factors to complicate the analysis.

References

[1] Al-Emadi, S., Al-Ali, A., & Al-Ali, A. (2021). Audio-based drone detection and identification using deep learning techniques with dataset enhancement through generative adversarial networks. *Sensors, 21*(15). https://doi.org/10.3390/s21154953

[2] Al-IQubaydhi, N., Alenezi, A., Alanazi, T., Senyor, A., Alanezi, N., Alotaibi, B., Alotaibi, M., Razaque, A., & Hariri, S. (2024). Deep learning for unmanned aerial vehicles detection: A review. *Computer Science Review, 51,* 100614. https://doi.org/10.1016/j.cosrev.2023.100614

[3] Aydın, I., & Kızılay, E. (2022). Development of a new light-weight convolutional neural network for acoustic-based amateur drone detection. *Applied Acoustics*, *193*, 108773. https://doi.org/10.1016/j.apacoust.2022.108773

[4] Basak, S., Rajendran, S., Pollin, S., & Scheers, B. (2021, January). Drone classification from RF fingerprints using deep residual nets. In *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)* (pp. 548-555). IEEE. https://doi.org/10.1109/COMSNETS51098.2021.9352891

[5] Biswas, D., & Tiwari, A. (2024). Utilizing Computer Vision and Deep Learning to Detect and Monitor Insects in Real Time by Analyzing Camera Trap Images. *Natural and Engineering Sciences, 9*(2), 280-292. https://doi.org/10.28978/nesciences.1575480

[6] Casabianca, P., & Zhang, Y. (2021). Acoustic-based UAV detection using late fusion of deep neural networks. *Drones*, *5*(3), 54. https://doi.org/10.3390/drones5030054

[7] Cheng, Y., Liu, X., Zhang, L., & Zhao, M. (2022). A deep CNN approach for acoustic classification of UAVs based on spectrogram analysis. *Journal of Intelligent & Robotic Systems, 104*(1), 67–78. https://doi.org/10.1007/s10846-021-01435-w

[8] Daniels, J. P. (2018). Venezuela's Nicolás Maduro survives apparent assassination attempt. *The Guardian*, *5*.

[9] Dar, D., Lacombe, L., Feist, J., & Maitra, N. T. (2022). Atom.

[10] Das, J. K., Ghosh, A., Pal, A. K., Dutta, S., & Chakrabarty, A. (2020, October). Urban sound classification using convolutional neural network and long short term memory based on multiple features. In *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)* (pp. 1-9). IEEE. https://doi.org/10.1109/ICDS50568.2020.9268723

[11] Demir, F., Abdullah, D. A., & Sengur, A. (2020). A new deep CNN model for environmental sound classification. *IEEE Access, 8,* 66529–66537. https://doi.org/10.1109/ACCESS.2020.2984903

[12] Fang, M., Tan, Z., Tang, Y., Chen, W., Huang, H., Dananjayan, S., ... & Luo, S. (2024). Pest-ConFormer: A hybrid CNN-Transformer architecture for large-scale multi-class crop pest recognition. *Expert Systems with Applications*, *255*, 124833. https://doi.org/10.1016/j.eswa.2024.124833

[13] Frid, A., Ben-Shimol, Y., Manor, E., & Greenberg, S. (2024). Drones detection using a fusion of RF and acoustic features and deep neural networks. *Sensors*, *24*(8), 2427. https://doi.org/10.3390/s24082427

[14] Ganesh, R., & Siva Kumar, R. (2021). Diagnosis of Brain Tumor Using Artificial Neural Network. *International Academic Journal of Innovative Research, 8*(1), 06–10. https://doi.org/10.9756/IAJIR/V8I1/IAJIR0802

[15] Gong, Y., Chung, Y. A., & Glass, J. (2021). Ast: Audio spectrogram transformer.

[16] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition.

[17] Hendria, W. F., Phan, Q. T., Adzaka, F., & Jeong, C. (2023). Combining transformer and CNN for object detection in UAV imagery. *ICT Express*, *9*(2), 258-263. https://doi.org/10.1016/j.icte.2021.12.006

[18] Khan, M. A., Menouar, H., Eldeeb, A., Abu-Dayya, A., & Salim, F. D. (2022). On the detection of unauthorized drones—Techniques and future perspectives: A review. *IEEE Sensors Journal, 22*(12), 11439–11455.

[19] Knoedler, B., Zemmari, R., & Koch, W. (2016, May). On the detection of small UAV using a GSM passive coherent location system. In *2016 17th International Radar Symposium (IRS)* (pp. 1-4). IEEE. https://doi.org/10.1109/IRS.2016.7497375

[20] Kumar, R., & Singh, P. (2021). Drone sound classification using shallow CNNs for surveillance applications. *International Journal of Applied Engineering Research, 16*(9), 733–740.

[21] Kümmritz, S. (2024). The sound of surveillance: Enhancing machine learning-driven drone detection with advanced acoustic augmentation. *Drones*, *8*(3), 105. https://doi.org/10.3390/drones8030105

[22] Lee, S., Han, J., & Kim, K. (2021). A hybrid CNN-LSTM network for UAV sound classification under noisy conditions. *IEEE Access, 9,* 14256–14265.

[23] Li, Z., Zhang, X., Wu, Y., & Yang, J. (2022). Acoustic UAV detection using MFCC features and ensemble SVM classifiers.

[24] Liu, Y., Wan, X., Tang, H., Yi, J., Cheng, Y., & Zhang, X. (2017, May). Digital television based passive bistatic radar system for drone detection. In *2017 IEEE Radar Conference (RadarConf)* (pp. 1493-1497). IEEE.

[25] Maaroof, M. K. A., & Bouhlel, M. S. (2025). Real-Time Object Detection Using YOLO-8 Model: A Drone-Based Approach. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 16*(1), 190-204. https://doi.org/10.58346/JOWUA.2025.I1.011

[26] Mohandas, R., Veena, S., Kirubasri, G., Mary, I. T. B., & Udayakumar, R. (2024). Federated Learning with Homomorphic Encryption for Ensuring Privacy in Medical Data. *Indian Journal of Information Sources and Services, 14*(2), 17-23.

[27] Nejad, N. D. (2015). Diagnosis of heart disease and hyperacidity of stomach through iridology based on the neural network. *International Academic Journal of Science and Engineering, 2*(1), 120–128.

[28] Ortega, R., & Pérez, J. (2020). Noise-robust audio surveillance for unmanned aircraft detection.

[29] Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 206-219. https://doi.org/10.1109/JSTSP.2019.2908700

[30] Rahman, M. H., Sejan, M. A. S., Aziz, M. A., Tabassum, R., Baik, J. I., & Song, H. K. (2024). A comprehensive survey of unmanned aerial vehicles detection and classification using machine learning approach: Challenges, solutions, and future directions. *Remote Sensing*, *16*(5), 879. https://doi.org/10.3390/rs16050879

[31] Santhi, P., Deeban, N., Jeyapunitha, N., Muthukumaran, B., & Ravikumar, R. (2020). Prediction of Diabetes Using Neural Networks. *International Journal of Advances in Engineering and Emerging Technology, 11*(2), 60–68.

[32] Shi, Z., Chang, X., Yang, C., Wu, Z., & Wu, J. (2020). An acoustic-based surveillance system for amateur drones detection and localization. *IEEE transactions on vehicular technology*, *69*(3), 2731-2739. https://doi.org/10.1109/TVT.2020.2964110

[33] Sricharan, K., & Venkat, M. (2023). Real-time drone detection using deep learning (Vol. 1). *Atlantis Press International BV.* https://doi.org/10.2991/978-94-6463-252-1_91

[34] Sudarsanam, P., Politis, A., & Drossos, K. (2021). Assessment of self-attention on learned features for sound event localization and detection. http://arxiv.org/abs/2107.09388

[35] Tsalera, E., Papadakis, A., & Samarakou, M. (2021). Comparison of pre-trained CNNs for audio classification using transfer learning. *Journal of Sensor and Actuator Networks, 10*(4). https://doi.org/10.3390/jsan10040072
[36] Utebayeva, D., Ilipbayeva, L., & Matson, E. T. (2022). Practical study of recurrent neural networks for efficient real-time drone sound detection: A review. *Drones*, *7*(1), 26. https://doi.org/10.3390/drones7010026
[37] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
[38] Wang, M., Yao, Y., Qiu, H., & Song, X. (2022). Adaptive memory-controlled self-attention for polyphonic sound event detection. *Symmetry*, *14*(2), 366. https://doi.org/10.3390/sym14020366
[39] Yan, X., Han, B., Su, Z., & Hao, J. (2023). SignalFormer: Hybrid Transformer for Automatic Drone Identification Based on Drone RF Signals. *Sensors*, *23*(22), 9098. https://doi.org/10.3390/s23229098
[40] Zhang, Y., Chen, L., & Li, F. (2021). Deep residual learning for UAV audio recognition in urban environments. *Journal of Aerospace Computing, 18*(7), 1123–1134. https://doi.org/10.2514/1.I010856
[41] Zhao, Y., & Lin, F. (2023). Audio event classification using transformer architectures in noisy environments. *IEEE Transactions on Multimedia.*

Authors Biography

Shahad Waleed Jassim is a Ph.D. student in the Department of Electrical Engineering, College of Engineering, at the University of Babylon, Iraq. She holds a Master's degree in Electronics and Communications Engineering from the same institution. Her research interests include a wide range of topics in electronics and telecommunications, particularly in the areas of signal processing, embedded systems, and wireless communication.

Saad Saffah Hreshee is a professor in the Department of Electrical Engineering at the College of Engineering, University of Babylon, Iraq. He received his professor Dr. in Electronic and Communication Engineering and has extensive academic and research experience in wireless communication, embedded systems, and digital signal processing. He has supervised numerous graduate students and contributed to several national and international research projects. His research interests include wireless sensor networks, IoT systems, and AI applications in communication systems.