From Sound to Sight: Audio-Visual Fusion and Deep Learning for Drone Detection

Ildi Alla, Hervé B Olou, Valeria Loscri, Marco Levorato

From Sound to Sight: Audio-Visual Fusion and Deep Learning for Drone Detection

Ildi Alla
Inria Lille-Nord Europe
Lille, France
ildi.alla@inria.fr

Hervé B. Olou
University of Abomey-Calavi
Porto-Novo, Benin
herve.olou@imsp-uac.org

Valeria Loscri
Inria Lille-Nord Europe
Lille, France
valeria.loscri@inria.fr

Marco Levorato
University of California
Irvine, United States
levorato@uci.edu

ABSTRACT

The proliferation of airborne drones, while instrumental to a broad range of applications, has led to an increased number of regulatory non-compliance incidents. The ubiquitous unmanned aerial vehicles (UAVs) are posing security risks, since they have started to be used for cybercrimes. Effective detection of illicit drones in restricted areas is paramount. Evolved drones are more and more sophisticated, and sometimes they do not emit RF-based signals, making inapplicable RF-based detection solutions. Different from existing work, this paper introduces a neural sensor fusion framework for drone detection based on both audio and video data to accurately identify drones and differentiate them from similar objects at long distances. Our design adopts a late fusion approach using the Weighted Average and Random Forest algorithm on the visual and auditory classification pipeline. Specifically, we process infrared data using a You Only Look Once (YOLO) v5 model due to its balance between inference time and accuracy. For the audio stream, we evaluate Long Short-Term Memory (LSTM) and Convolutional Recurrent Neural Network (CRNN) models and demonstrate the superiority of the CRNN model through Mel-Frequency Cepstral Coefficients (MFCC) features. To demonstrate the robustness of our audio-visual fusion approach, we validate it in extensive scenarios, with impaired audio/video data. Our results demonstrate that multimodal fusion significantly improves drone detection, outperforming traditional single-modality systems in complex environments. Additionally, our system provides predictions rapidly, in just 0.382 seconds, making it well-suited for real-time applications.

CCS CONCEPTS

• **Networks** → **Wireless access networks**; • **Security and privacy** → **Intrusion/anomaly detection and malware mitigation**; • **Computing methodologies** → **Neural networks**; **Ensemble methods**.

KEYWORDS

drone detection, deep learning, audio classification, infrared imaging, late fusion.

1 INTRODUCTION

The increasing use of airborne drones in commercial applications such as surveillance, logistics, and photography raises concerns about maintaining a safe and private airspace. The need for robust drone detection mechanisms is increased by the possibility of incidents such as aircraft interference, invasive monitoring, and threats to essential services [39]. Some threats related to drones encompass illegal surveillance or also terroristic attacks, as occurred in [8].

The detection of drones, especially during nocturnal hours, is then of paramount importance due to the reduced visibility and increased likelihood of covert operations. Some of the possible drone detection approaches are based on humming sounds from drones (audio), Radio Frequency (RF) signals from the drone controller, etc. Infrared (IR) sensors have been recognized as a pivotal technology for enhancing nocturnal detection capabilities, utilizing thermal signatures to detect drones in conditions where traditional visual spectrum sensors fail [33]. However, existing systems relying solely on IR imaging face limitations in terms of detection accuracy and environmental adaptability. On the other hand, audio-based detection solutions also offer advantages under certain conditions, but their effectiveness can be compromised in noisy environments or be affected by factors such as wind direction and temperature.

This paper aims to bridge this gap by proposing a novel drone detection system that synergies the strengths of both IR imaging and acoustic sensing. We posit that integrating IR sensors' thermal detection capabilities with the light-independent drone identification offered by acoustic sensors can significantly enhance detection efficiency [5, 34]. This integration is achieved through late sensor fusion, a strategy shown to augment detection robustness by leveraging the complementary features of each sensing modality, thereby reducing faults and improving reliability in diverse operational scenarios [33, 34]. This work's contributions can be summarized as follows:

- We extend the application of Long Short-Term Memory (LSTM) and Convolutional Recurrent Neural Network (CRNN) models to drone audio recognition. Our approach includes novel customization of the LSTM architecture, optimizing it for Mel-Frequency Cepstral Coefficients (MFCC) feature processing. Additionally, we integrate a specialized CRNN model in PyTorch for Log-Mel spectrogram and MFCC feature analysis, leading to improved audio recognition performance.
- We have developed a preprocessing technique that is not only tailored to the unique characteristics of IR imagery but also exhibits versatility for adaptation to various datasets.

Our methodology adapts You Only Look Once (YOLO) models to IR imagery, employing layer freezing to enhance computational efficiency. We conduct a comparative analysis of YOLOv3 and YOLOv5 models on IR datasets, a perspective that is largely unexplored in the existing literature, to the best of our knowledge.

- We propose a novel fusion approach, termed Dynamic Weighted Average Fusion (DWAF). In particular, DWAF is based on the combined outputs of the CRNN and YOLOv5. Specific adaptive weights are assigned to each model based on its performance during the prediction process. A further stage, called Multi-Modal Tree Fusion (MMTF) is applied to the predictions, based on a trained classifier, to enhance the accuracy of the prediction.

The evaluation outcomes demonstrate the superior performance of our multi-modal approach, which significantly enhances accuracy even in scenarios where image and audio data may lack precision. Our late fusion strategies yield high accuracy rates; specifically, the MMTF technique achieves an accuracy of 96.02%, while the DWAF method reaches 89.31% accuracy, both methods notably reducing false positive rates. The MMTF technique demonstrates its suitability for real-time detection and tracking by performing the detection in only 0.382 seconds.

The rest of the paper is organized as follows. Section 2 reviews existing literature on drone detection. Based on the limitations identified in these works, we propose the Threat Model in Section 3. Section 4 introduces the proposed drone detection system. Section 5 presents and discusses the results obtained. Finally, Section 6 closes with the conclusion and future work.

2 RELATED WORK

In the critical domain of airspace security, Counter-Unmanned Aerial Vehicle (C-UAV) systems, or anti-drone systems, are pivotal for neutralizing the threat of unauthorized drones. These systems, which include a range of technologies from lasers to microwaves, are essential for protecting sensitive civilian areas from potential aerial intrusions [24]. Existing research in C-UAV systems primarily focuses on integrating various detection and neutralization technologies, underscoring the necessity for adaptable and cost-effective strategies to counter increasingly sophisticated aerial threats [35, 41]. The suite of anti-drone technologies features radar, acoustic sensors, and infrared systems, often used in synergy to create a robust defense mechanism. The success of these systems relies on the critical first step of precise drone detection, which underpins the entire counteraction process. While comprehensive, the current approaches often fall short in addressing key challenges such as scalability, environmental adaptability, and multi-drone detection capabilities [12].

Single Sensor Detection. The landscape of drone detection has evolved from reliance on single-sensor systems, such as RF, acoustic, camera, or radar technologies, to address the increasing complexity of Unmanned Aerial Vehicle (UAV) identification. Despite the advancements, these methods face significant challenges, including high false-positive rates and environmental limitations, underscoring the need for improved detection strategies [21, 33]. A notable effort in RF-based detection utilized a comprehensive drone RF signal database to fuel three Deep Neural Networks (DNNs) for identifying drone presence, type, and flight mode. However, the performance noticeably diminished as the classification complexity rose, pointing to the inherent difficulty of distinguishing between drones from the same manufacturer due to RF spectrum similarities [2]. This observation suggests an urgent demand for advanced algorithms or more sophisticated network architectures to enhance classification precision. Furthermore, existing designs often overlook the variety of drone types and operational ranges, thereby restricting their detection efficacy across different protocols and frequencies, and struggle to accurately identify multiple drones in densely populated drone environments [23].

Visual and thermal identifications through cameras present a direct approach to drone detection. Studies highlighting the efficacy of C-LBP, C-HAAR, and C-HOG methods underscore their potential in precise detection and distance estimation, though they are hampered by non-optimized processing in real-time applications, suggesting an avenue for computational optimization [10]. IR cameras extend the detection capabilities by leveraging thermal signatures, particularly effective in low visibility conditions, albeit with diminished utility at extended ranges [26]. Furthermore, the implementation of YOLOv4-based systems marks a pivotal enhancement in detection accuracy and speed, with the caveat of reduced efficacy at high altitudes, emphasizing the need for algorithmic innovation and expanded datasets [31]. The exploration of background subtraction methods, integrated with Fourier descriptors and HOG features for Support Vector Machine (SVM) classification, reveals potential, though challenges remain in accurately distinguishing drones from birds in complex environments [40]. Additionally, a study introducing a real-time drone detection algorithm that combines moving object detection with CNN-based classification demonstrates impressive processing speed, yet faces limitations in dynamic backgrounds, impacting detection precision and increasing false positives and negatives [28].

The domain of drone detection has seen advancements through various methodologies, particularly in audio analysis. The authors in [17] presented a system with promising indoor performance that, however, struggles with outdoor false alarms. The study in [20] explored correlation methods for sound detection, pointing out the need for improvements against environmental noise and distance challenges. Machine learning advancements were highlighted in [11], where a Recurrent Neural Network (RNN) demonstrated accuracy in detecting drones up to 150 meters, albeit limited to binary classification. The comparison between Plotted Image Machine Learning (PIL) and K-Nearest Neighbors (KNN) in [15] revealed PIL's superior accuracy, emphasizing the importance of extensive data. Lastly, the use of Hidden Markov Models (HMM) for UAV detection was explored in [30], which faced limitations due to the scarcity of training data.

Sensor Fusion Detection. Recent research in sensor fusion strategies has shown promising advancements in enhancing the robustness of object detection systems [29]. These strategies, encompassing early, intermediate, and late fusion phases, leverage the strengths of diverse sensor inputs to improve detection accuracy and reliability. Early fusion, as explored in [16], combines raw camera and radar data for superior 3D object detection. Intermediate fusion, demonstrated by the innovative work of integrating DNNs

and CNNs for drone detection, faces challenges in data processing complexity, despite achieving a validation accuracy of 75% [3]. Late fusion techniques, which evaluate sensor data independently before combining outcomes, have been shown to significantly reduce false detections, as evidenced by multi-sensor drone detection systems achieving notable F1-scores through the integration of video, thermal infrared, and audio sensors [33]. Furthermore, the fusion of audio and vision features using CNNs and SVMs for audio and YOLOv5 for visual data has been highlighted for its potential to increase detection accuracy across various distances, despite limitations in drone-type diversity and the call for further exploration into real-time detection systems and deep learning methods [13, 14].

Despite these advancements, a significant gap remains in developing a comprehensive and scalable C-UAV system capable of accurate, real-time detection in diverse and challenging environments, particularly in scenarios involving multiple drones and complex backgrounds. This research aims to address these gaps by focusing on late fusion strategies. We propose to enhance the robustness and applicability of late fusion in drone detection systems, particularly for real-time applications, by extending its capabilities to handle a broader range of environmental conditions and operational distances. This approach seeks to capitalize on the inherent flexibility and adaptability of late fusion, aiming to overcome the limitations observed in current technologies and methodologies.

3 THREAT MODEL

The Threat Model considered in this work is illustrated in Figure 1. Our system considers a large set of malicious drones, where nodes can overfly a restricted area, but are not necessarily emitting RF signals. This approach enables the detection of drones that might evade traditional RF-based systems, significantly broadening our system's applicability.

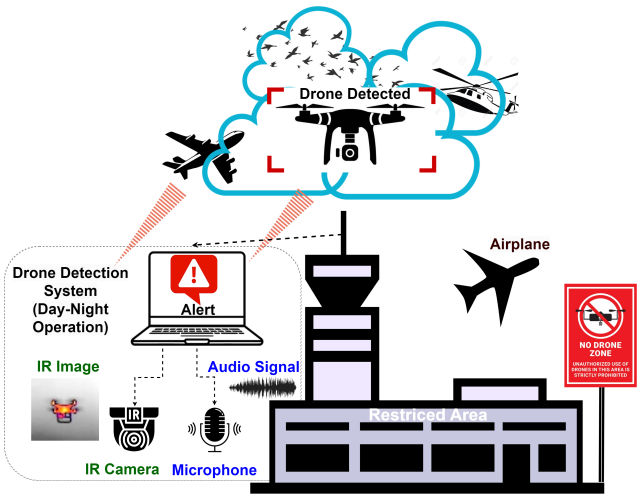

Figure 1: Threat Detection Model.

In our model, we consider a detector node installed on the ground and collecting both audio and video data. The position of the detector is fixed. The model developed to run in the detector, is built to handle various scenarios, noisy and low-noisy, and can work with both audio and video data, and also with just one source of data. This flexibility ensures robust detection capabilities across diverse operational contexts without the need for prior knowledge about the drones, enhancing the system's adaptability and responsiveness.

Our system is designed to detect drones infiltrating protected spaces undetected by traditional means. By harnessing both audio and IR imaging, we significantly reduce the likelihood of evasion, ensuring a high degree of detection.

The setup considered in this work is based on a laptop running the intelligence based on the fusion approaches, taking in input the audio and IR data. The initial setup consists of a single detector, to demonstrate the feasibility of the data fusion approach and its high accuracy also in adversarial conditions (e.g., with high noise, scarce visibility, etc.). This system can be easily extended to include more detectors opportunistically deployed and communicate with each other to exchange data from different viewpoints.

4 PROPOSED FRAMEWORK

In Figure 2, we present our drone detection system, which integrates audio and infrared video modalities to leverage their unique detection capabilities. Audio is essential for its ability to capture acoustic signatures, which are particularly valuable in low-visibility environments. In contrast, infrared videos are indispensable for detecting thermal signatures, enabling reliable detection even in visually challenging conditions. Our system is meticulously trained on four distinct classes — airplanes, birds, drones, and helicopters — to minimize potential misidentifications that could arise from distant or overlapping signatures. Following detection, a fusion strategy synergizes the insights from both modalities to ensure real-time and accurate drone identification among other aerial entities.

The framework is comprehensively detailed in the following subsections, beginning with the datasets employed, progressing through the deep learning models utilized, and culminating with an exposition of the advanced late fusion techniques.

4.1 Datasets

The datasets include data captured using a thermal IR camera and a microphone. The audio tracks in the datasets are either extracted from the videos or recorded separately under different conditions.

The audio data sourced from [32] comprises 90 files in WAV format with a sampling frequency of 44.1 KHz. The authors provide data for three classes: *Background*, *Drone*, and *Helicopter*. To ensure the same number of classes as in the infrared dataset, we incorporated data for airplanes and birds from [1] to enhance the comprehensiveness of the dataset. In total, we have 130 audio files, each lasting 10 seconds. Figure 3 showcases a spectrogram of a drone, distinguished by a red line that signifies the drone's unique frequency signature—a feature not present in the spectrograms of other audio samples, which exhibit a mix of low and high frequencies.

We examined the IR Video dataset [32], comprising 365 IR videos captured from various distances, up to a maximum of 200m. Each

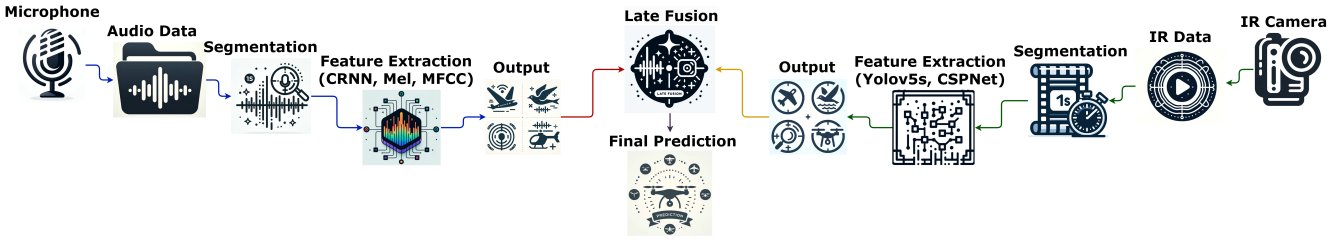

Figure 2: Workflow of the drone detection system.

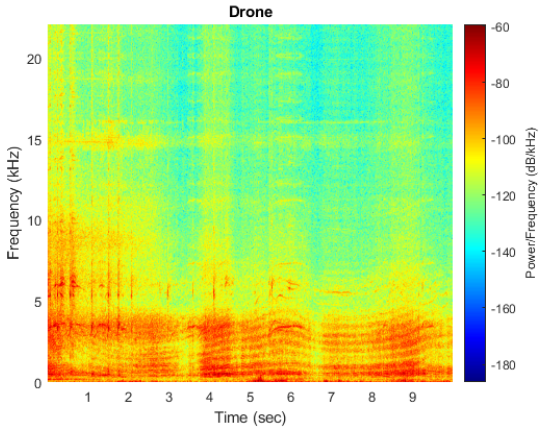

Figure 3: Spectrogram of drone audio sample.

video is ten seconds long and has a resolution of 320 × 256 pixels. From 200 videos, we extracted frames at a rate of one every 0.1 second, resulting in approximately 20,688 annotated images. The annotations categorize the subjects into four classes: *Airplane*, *Bird*, *Drone*, and *Helicopter*, as shown in Figure 4. The weather conditions in the dataset range from clear and sunny to scattered clouds and complete overcast.

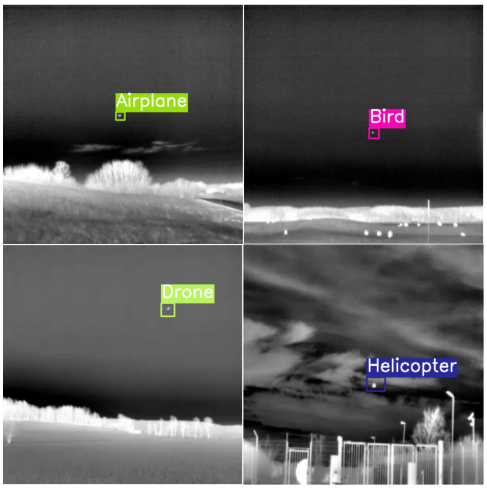

Figure 4: Dataset IR images.

4.2 LSTM and CRNN models

4.2.1 Models description. Among various approaches, LSTM networks [38] excel in capturing the temporal dynamics of drone audio. Conversely, CRNNs [25] combine temporal sequence learning with superior feature extraction capabilities. This synergy makes CRNNs robust choices for the complex task of drone audio detection, particularly in noisy and real-time environments. In the CRNN model architecture, detailed in Table 1, we configured the model with a learning rate of 0.0001 and employed the *Adam optimizer* in PyTorch, alongside a patience setting of 10. The training was

Table 1: CRNN Model Architecture

Configuration	Description
Convolution	16 kernels of 5x5 size, ReLU activation
Max Pooling	5x5 pooling size with stride 2
Batch Normalization	-
Reshape for LSTM	-
LSTM	32 memory units
Flatten	-
Dense	32 neurons, ReLU activation
Dropout	Dropout with a rate of 0.3
Dense (Output)	4 neurons, Softmax activation

conducted over approximately 30 epochs, incorporating strategies such as early stopping and model checkpoints to preserve the best-performing model and minimize the risk of overfitting. A batch size of 16 was found to yield optimal results after evaluating various sizes [6].

The LSTM model, implemented in Keras, utilizes Log-Mel spectrograms or MFCC features for input. Its architecture features a combination of Layer Normalization, TimeDistributed Dense layers, and a Bidirectional LSTM layer with a skip connection, specifically designed to capture temporal features effectively. Additionally, the model incorporates regularization techniques and dropout to mitigate overfitting. We adjusted the feature extraction process and class specifications within the LSTM model to align with our specific research requirements, ensuring consistency in parameters with those used for the CRNN model.

4.2.2 Audio Feature Extraction. According to [6], segmenting audio into 1-second intervals has produced more favorable results than another method, which involves using shorter audio segments. Furthermore, as indicated by [9], this specific time interval is more

promising compared to other intervals. This is because it allows the deep learning algorithm to more precisely learn the features, as opposed to processing the entire audio in one go. To diversify our audio dataset, we experimented with various data augmentation techniques, including pitch shifting, noise injection, dynamic range compression, and reverberation. However, according to [22], pitch shifting and white noise were the only methods that positively impacted performance without adversely affecting detection classes. Building on these results, we applied pitch shifting and added white noise to the audio recordings, thereby enhancing the system's ability. The augmentation methods used are described as follows:

- *Pitch shifting*: The pitch of every audio signal within the datasets is elevated by a factor of +2.
- *White noise*: The audio segments are subjected to white noise with an intensity of 0.05.

Various feature extraction techniques are capable of capturing distinct attributes. In detection tasks, the Log-Mel spectrogram and MFCC are frequently employed as features [7, 11], with MFCCs capturing the audio's timbral aspects, while Log-Mel spectrograms are utilized for analyzing the signal's frequency content. Each sound sample was adjusted to a sampling rate of 22.05 kHz, and its distinctive features were extracted using Librosa [19], a Python package designed for audio analysis. Our study focuses on examining the behavior of both the Log-Mel spectrogram and MFCC features. Specifically, we analyzed the Mel spectrograms and MFCCs across segments, identifying 44 frames in each. The number of Fast Fourier Transform (FFT) points was set to 2048, with a hop length of 512, to ensure detailed and consistent feature extraction.

We conducted experiments to investigate the impact of various MFCC features and Mel scales on the results. Consequently, each audio segment is represented by a dimension of $N \times 44$, where N specifies either the number of Mel-frequency cepstral coefficients extracted from a segment or the number of Mel scale frequency bands. Upon fine-tuning this parameter, we established that the optimal input size for the Log-Mel spectrogram is 90×44, while for the MFCC features, it is 40×44. This methodology is illustrated in Figure 5.

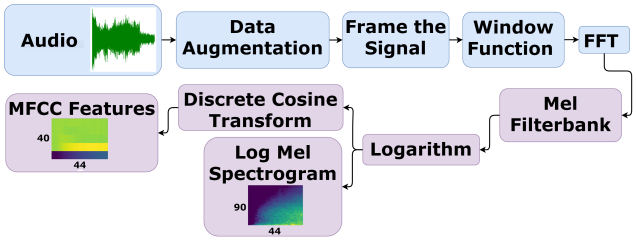

Figure 5: Feature extraction steps.

4.3 YOLO's models

4.3.1 Models description. In pursuit of a real-time object detection model suitable for constrained hardware, we first integrated YOLOv3-tiny into our framework. This simple version of YOLO, with its reduced layer count of 38, provides a compromise between speed and accuracy. Its lightweight design ensures faster inference times, essential for real-time applications, at the cost of some precision in detection, especially in multi-class scenarios for distant objects. Building upon the insights gained from YOLOv3-tiny, we then explored the full YOLOv3 architecture, which employs a robust Darknet-53 backbone. This comprehensive structure, extended to 190 layers including convolutional strides, bottleneck blocks, and multi-scale predictions, is referenced in [37]. Despite its higher precision, the complexity of YOLOv3 leads to increased inference times. After experimenting with different numbers of blocks to freeze, we decided to freeze 10 blocks in the models to achieve a balance between training time and model precision. This incremental approach, from the agility of YOLOv3-tiny to the precision of YOLOv3, sets the stage for our subsequent evaluation against YOLOv5, where we seek to benchmark performance and identify the optimal model configuration for real-time object detection in our specific use case.

4.3.2 IR Feature Extraction. Before training the model, we resized the images to dimensions of 414×416. The ground truth generated by the *Video Labeler* app [18] uses a specific format for bounding box positions. This format employs $x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$ in pixels. Conversely, YOLO label format uses x_{center}, y_{center}, $width$, and $height$. Consequently, we meticulously review and normalize the labels to align them with the YOLO format.

The YOLOv5 architecture is structured into three primary segments: the backbone, neck, and head [4]. The backbone incorporates the innovative Cross-Stage Partial Network (CSPNet), which addresses gradient-related challenges and reduces model complexity, leading to enhanced inference speed and accuracy in object detection. This section acts as a feature extractor, using multiple convolutional layers and a Spatial Pyramid Pooling (SPP) mechanism. The subsequent segment, the neck or the Path Aggregation Network (PANet) facilitates feature fusion by transmitting extracted features to deeper layers. The concluding segment, the head, is responsible for object detection, employing various convolutional layers to predict object classes, outline bounding boxes, and assign class confidence scores. Figure 6 shows the overall architecture of YOLOv5. Among the various YOLOv5 models [36], we selected YOLOv5s (small) and YOLOv5m (medium) as backbone architectures due to our hardware constraints, as the models are trained on our PC, necessitating a balance between computational efficiency and detection performance.

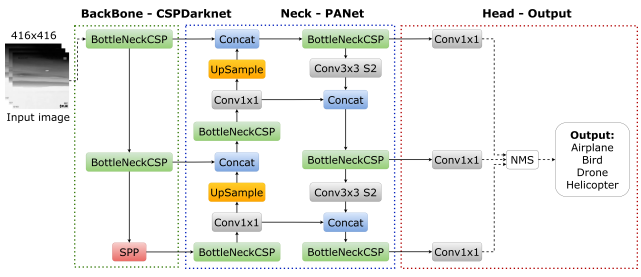

Figure 6: YOLOv5 architecture.

4.4 Fusion approach

Late fusion or decision-level fusion, is an approach used to combine predictions from multiple models or sensors. We chose late fusion for its suitability in handling heterogeneous data from multiple sources, allowing for the independent processing of diverse input and output formats. This approach ensures the preservation of unique sensor characteristics, enhances system flexibility for integrating new technologies, and improves robustness against sensor failures. We have implemented two different late fusion approaches as follows.

4.4.1 Dynamic Weighted Average Fusion. The system described employs a sophisticated technique that integrates probabilistic predictions from two distinct sources: audio and video, as detailed in *Algorithm 1* with respective abbreviations explained in Table 2. This integration begins with an unbiased starting point, where each modality's prediction is initially assigned an equal weight of 0.5. These modalities are then processed independently through specialized models to generate their own set of class probabilities. At the core of this integration strategy is an adaptive weighting process, facilitated by three key functions.

Table 2: Abbreviations Used in Pseudocode

Abbreviation	Description
a_p	Audio Probability
v_p	Video Probability
a_w, v_w	Audio, Video Weights
b_s	Batch Size
$WtAvgFus$	Weighted Average Fusion Function
$UpdWts$	Update Weights Function
$GetConf$	Get Confidence Function
wt_p	Weighted Probability
tot_w, tot_c	Total Weight, Total Confidence
f_p	Fused Probability
$pred_c$	Predicted Class
a_{pw}, v_{cw}	Audio, Video Probability Windows
$conf$	Confidences List
cb	Current Batch
a_c	Audio Confidence
v_c	Video Confidence

The '*GetConf*' function plays a crucial role by calculating the average confidence of each modality. It achieves this by extracting the maximum confidence value from the probability distributions of each segment, providing a quantitative basis for weight adjustment. Following this, the '*WtAvgFus*' function combines the predictions, adjusting weights dynamically to favor the more confident modality, thus enhancing predictive performance. This dynamic adjustment is operationalized through the '*UpdWts*' function, which recalibrates weights for the next batch of inputs based on the recent confidence levels of the modalities.

The methodology of processing inputs in batches, as defined by a b_s parameter, allows for the accumulation of data necessary for informed weight adjustments. This ensures that the final prediction not only leverages the strengths of both audio and video inputs

Algorithm 1 Dynamic Weighted Avgerage Fusion

Require: a_p, v_p, a_w, v_w, b_s
Ensure: Weighted average probability & predicted class
function WtAvgFus(a_p, v_p, a_w, v_w)
  if a_p = None $\vee$ v_p = None **then**
    Handle missing predictions
  end if
  $wt_p \leftarrow (a_p \cdot a_w) + (v_p \cdot v_w)$
  $tot_w \leftarrow a_w + v_w$
  return wt_p/tot_w **if** $tot_w > 0$ **else** wt_p
end function
function UpdWts(a_c, v_c)
  $tot_c \leftarrow a_c + v_c$
  return $(a_c/tot_c, v_c/tot_c)$ **if** $tot_c > 0$ **else** $(0.5, 0.5)$
end function
function GetConf($probs$)
  Calculate avg. confidence from model probs.
  Initialize $confidences \leftarrow []$
  for *chunk* in $probs$ **do**
    if *chunk* is not empty **then**
      Append max(*chunk*) to $conf$
    end if
  end for
  return mean($conf$) **if** $conf$ not empty **else** 0
end function
Init. a_{pw}, v_{pw} with b_s
for $\forall segment$ **do**
  $f_p \leftarrow$ WtAvgFus(a_p, v_p, a_w, v_w)
  $pred_c \leftarrow$ argmax(f_p)
  $cb = cb + 1$
  if $cb \mod b_s = 0$ **then**
    $a_c, v_c \leftarrow$ GetConf(a_{pw}, v_{pw})
    $a_w, v_w \leftarrow$ UpdWts(a_c, v_c)
  end if
end for

but also dynamically balances their contributions to optimize overall accuracy. The adaptive, confidence-informed fusion approach enables the system to achieve robust performance across varying conditions, efficiently handling fluctuations in the reliability of individual modalities.

Furthermore, we tested an alternate technique that updates weights based on the accuracies of each modality's predictions relative to ground truths. Unlike the primary method, this technique does not utilize the '*GetConf*' function and omits the application of a sliding window mechanism to maintain recent confidences. Although this method promises higher results, it is limited by the availability of ground truths post-deployment, rendering it less feasible for real-world applications. This contrast highlights the system's innovative approach to optimizing prediction accuracy through adaptive weighting, addressing the challenges of integrating multimodal data sources.

4.4.2 Multi-Modal Tree Fusion. In our work, we introduce a novel multi-modal fusion technique designed to enhance classification performance by integrating audio and video data, as depicted in

Figure 7. This method excels in situations where predictions from audio or video sources are either missing or unreliable. By employing prediction labels and probabilities from the CRNN for audio and YOLOv5s for video, we create a robust framework for data analysis. To tackle the challenge of missing modality data, we impute such gaps with predetermined placeholder values, ensuring uniform input dimensions for our fusion process.

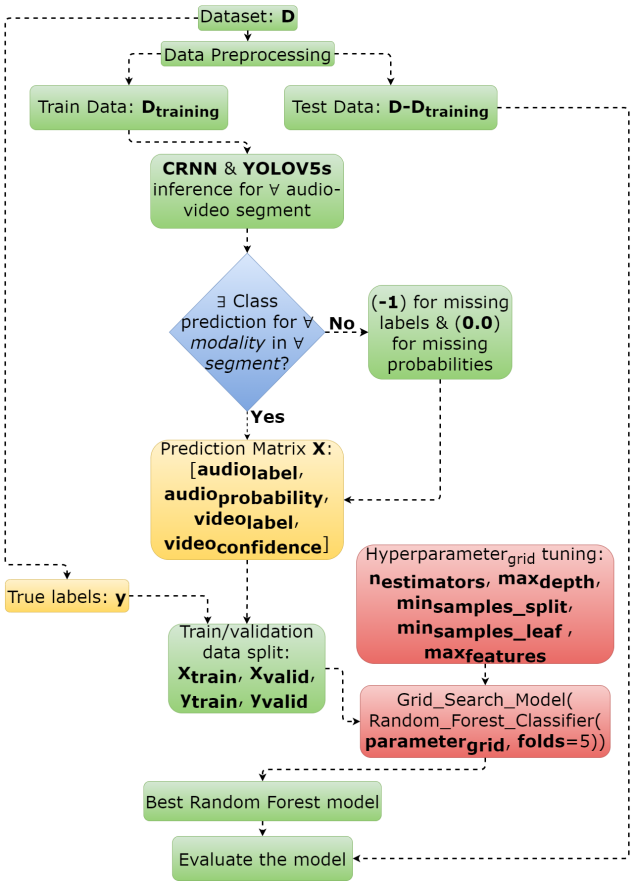

Figure 7: Multi-Modal Tree Fusion training procedure.

Central to our approach is the deployment of a Random Forest classifier, fine-tuned through GridSearchCV to identify an optimal configuration of 100 trees ($n_{estimators}$), a decision supported by 5-fold cross-validation. This meticulous optimization process guarantees that our fusion method is not only accurate but also resilient, capable of handling the intricacies of real-world data scenarios where incomplete or inconsistent data is common.

Our multi-modal tree fusion technique stands as a significant advancement in the field, demonstrating the ability to seamlessly synthesize insights from both auditory and visual cues, thereby significantly improving classification outcomes even in the face of challenging data conditions. This innovative approach promises to be a valuable asset in applications requiring nuanced data analysis and interpretation.

5 RESULTS AND DISCUSSION

5.1 Hardware specification

The system operates on Windows 11 with a 12th Gen Intel® Core™ i7-12800H processor, which has 14 cores and features a base clock speed of approximately 2.4 GHz. The system is complemented by 32 GB of RAM. It includes an NVIDIA RTX A1000 GPU with 4 GB of dedicated memory.

5.2 Drone sound recognition

From Table 3, we have divided the audio dataset into training, validation, and testing subsets, comprising 71%, 13%, and 16% of the dataset, respectively. To enhance the robustness of the model, techniques such as pitch shifting and noise injection have been applied to these subsets. Comparative analysis reveals that the

Table 3: Data Description

Type of data	Class	Audio	Image	Audio (s)	Video (s)
Train	airplane	64	4301	2690	2920
	bird	64	3616		
	drone	77	4334		
	helicopter	64	4299		
Validation	airplane	11	537	460	370
	bird	11	452		
	drone	13	541		
	helicopter	11	537		
Test	airplane	15	538	600	370
	bird	15	454		
	drone	15	542		
	helicopter	15	537		

CRNN model outperforms the LSTM model in audio classification tasks. This superiority is particularly evident in tasks involving the classification of multiple classes using MFCC features, as detailed in Table 4. The effectiveness of the CRNN model is underscored by its

Table 4: Evaluation Metrics for LSTM and CRNN Models

Model	Feature	Classes	Metric	
			CPU-time (ms)	Accuracy (%)
LSTM	Mel	Two	15.987	89
		Four	17.163	76
	MFCC	Two	48.861	92
		Four	44.134	82
CRNN	Mel	Two	12.257	95
		Four	13.007	82
	MFCC	Two	16.058	94
		Four	13.463	85

higher accuracy, a metric that quantifies the proportion of correctly identified predictions across all classes. Accuracy is calculated using the formula 1, where TP, FP, TN, and FN represent the counts of true positives, false positives, true negatives, and false negatives, respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In binary audio classification, the CRNN model efficiently differentiated "*drone*" from "*no drone*" sounds with superior accuracy and lower inference time over LSTM. The model's simplicity negates the need for GPU use, as it performs equally well on the CPU. Based on its performance, we proceed with the CRNN model using MFCC features.

5.3 Drone IR image detection

Due to GPU memory constraints, we opted to train YOLOv3 with a batch size of 4, a decision informed by its computational intensity, evidenced by its 154.6 Giga Floating Point Operations per Second (GFLOPs) demand. Despite this, YOLOv3's accuracy improvements plateaued after 25 epochs. In contrast, YOLOv5m, with a significantly lower computational requirement of only 47.9 GFLOPs, allowed for a larger batch size of 8 and exhibited continuous improvements in mean Average Precision (mAP) up to the 50th epoch. The performance of YOLO models is quantitatively assessed through metrics such as mAP@0.5 and mAP@0.5:0.95, which evaluate object detection capabilities across varying Intersections over Union (IoU) thresholds. Specifically, mAP@0.5 focuses on precision and recall at an IoU threshold of 0.5, whereas mAP@0.5:0.95 provides a more comprehensive assessment by averaging these metrics across IoU thresholds from 0.5 to 0.95.

To further elucidate the models' performance, we examined additional metrics:

1. Precision: The ratio of true positive predictions to total positive predictions, defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

2. Recall (Sensitivity): The ability of the model to identify all positive samples, calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

A balanced trade-off between 'Precision' and 'Recall' is crucial for an effective model. High precision indicates reliable detections, whereas high recall reflects the model's capability to identify most objects, albeit potentially at the cost of accuracy.

Our evaluation utilized a dataset of 2,071 infrared (IR) images. The YOLOv3-tiny variant, while achieving high frame rates suitable for real-time applications, exhibited limitations in accuracy, especially for small or distant objects. YOLOv3's training duration spanned 20.186 hours on a GPU, with performance diminishing over greater distances and with increased object classes. On the other hand, YOLOv5 models demonstrated variable outcomes. The YOLOv5s model, completing its training in approximately 5.189 hours, achieved high mAP values at a 50% IoU threshold but showed slight decreases in precision at stricter IoU thresholds and over longer distances. The YOLOv5m model, after around 10.031 hours of training, displayed superior precision and accuracy, albeit at the cost of longer inference times. Notably, freezing the first 10 layers of YOLOv5m reduced training time but also resulted in lower mAP scores and compromised detection reliability, paralleling our observations with YOLOv3.

The YOLOv5s model exhibits the capability to detect multiple objects simultaneously, a feature illustrated in Figure 8, overcoming the constraints observed in existing literature where alternative models struggle with the detection of multiple drones or distant objects. As summarized in Table 5, our analysis underscores YOLOv5s's advantageous balance of rapid training and inference speeds against YOLOv5m's marginally superior accuracy. Despite the efficiency gains from layer freezing, the consequent accuracy reduction led to our preference for YOLOv5s, considering its optimal performance and speed trade-off.

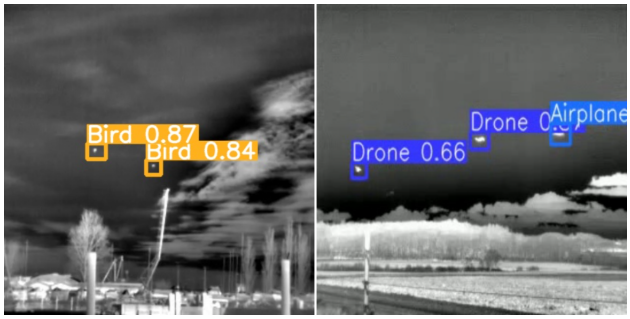

Figure 8: Multi-object detections.

Table 5: Detection Performance Metrics

Model	Precision	Recall	mAP50	mAP50-95	Inference (ms)	
					CPU	GPU
YOLOv3-tiny	0.985	0.983	0.988	0.685	84.1	6.7
YOLOv3 full training	0.995	0.995	0.994	0.807	688.9	46.1
YOLOv3 frozen layers	0.875	0.882	0.895	0.412	701.1	46.8
YOLOv5s	0.995	0.994	0.994	0.815	157.9	9.3
YOLOv5m full training	0.997	0.995	0.995	0.847	300.3	20.1
YOLOV5m frozen layers	0.987	0.988	0.99	0.728	302.6	20.2

5.4 Late fusion

Our study improves drone detection by employing late fusion techniques on 60 synchronized audio and video files in a combined dataset. This dataset included challenging cases from [32] for objects at distances up to 200 meters. To simulate the effects of distance, audio adjustments were made using the Python package PyDub [27]. The data, derived from various scenarios, posed challenges including low resolution, occlusion, the presence of multiple objects, and noisy audio conditions. These challenges were further compounded by diverse environmental noise and white noise introduced during augmentation. The fusion methods were specifically designed to overcome sensor performance limitations at extended distances, in noisy conditions, and in instances of missing data from predictions. Occasionally, segments lacking predictions are appended to the end of the files. This approach is undertaken to observe the resultant effects on the system, thereby enabling an understanding of the system's resilience and its response to such conditions. In the following, we will detail the results achieved with our techniques and compare them with those from existing detection methods.

5.4.1 Dynamic Weighted Average Fusion (DWAF). Regarding the performance of the DWAF system, we analyzed it by testing it with various values of the hyperparameter 'b_s' across two scenarios. The first scenario involves situations with no missing data, where predictions are made by each model independently. This implies that for the audio sensor, the corresponding class is perceivable, and for the IR camera, the object is within the line of sight. However, such ideal conditions are not always present, especially with static sensors. For instance, an object may be in motion and out of the camera's field of view, or the audio sensor might fail to detect it. To address these challenges, we introduced a second scenario accounting for missing sensor data. In instances where predictions are absent, we assign a marked class (-1) to each segment. Additionally, we evaluated the system's performance in updating weights related to ground truth. This assessment aimed to gauge the stability and accuracy of our system and to understand its deviation from the ideal case of ground truth, which provides the actual classes and thereby facilitates more accurate weight updates.

The parameter 'b_s' influences how often our model updates, affecting the efficacy of the learning process, as shown in Table 6. The choice of b_s can significantly affect the behavior of the system.

Table 6: Effect of batch size on DWAF accuracy according to scenarios

Batch Size	Accuracy (%)			
	Without Data Missing		Data Missing	
	Ground Truth	Confidence Score	Ground Truth	Confidence Score
Off	87.07		86.59	
5	**92.48**	86.77	**91.42**	86.89
10	92.18	88.87	90.81	87.96
20	90.53	88.12	89.91	87.35
30	90.38	88.72	88.86	87.65
50	89.02	89.32	88.10	88.55
70	87.82	89.77	87.19	89.01
100	85.56	**90.23**	84.64	**89.31**

A smaller b_s leads to more frequent updates (potentially more responsiveness to changes), while a larger b_s provides more stability in the weights but less responsiveness. We have experimented with various values of this parameter but observed no improvement in the results beyond a parameter value of 100. When the dynamic change of weights is not utilized (i.e., b_s = off, and the weights are fixed at 0.5), we observe that the accuracy is lower compared to when the b_s is set to 100 for the *'Confidence Score'* case. Additionally, with the b_s set to 5, the 'Ground Truth' case achieved the best performance across 665 audio-video segments. However, in the 'Confidence Score' case, which is closer to real-world conditions, we achieved an accuracy of 89.31%. The number of false positives for the drone class is 10. This indicates that the system has reduced the occurrence of false detections.

5.4.2 Multi-Modal Tree Fusion (MMTF). For the training of the MMTF system, we used 75% (489 samples) of the audio-video files for training purposes and the remaining 25% (176 samples) for testing. Similar to the DWAF approach, we evaluated the system's robustness in various scenarios by testing it both with and without missing data. To enrich our evaluation, beyond the metrics previously employed, we incorporated the F1-score—an indicator that provides a balanced measure of precision and recall by computing their harmonic mean as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

As evidenced by Table 7, the F1-scores and other metrics exhibit minimal variation, even in scenarios with missing data predictions. This highlights the MMTF system's stability and its ability to maintain high accuracy despite data gaps. Upon refining the system through hyperparameter tuning, we identified an optimal parameter set for the second scenario, characterized by: $max_{depth} = 10$, $max_{features} = sqrt$, $min_{samples\_leaf} = 1$, $min_{samples\_split} = 5$, and $n_{estimators} = 100$. This configuration demonstrated a marked improvement in classification accuracy over the DWAF approach, particularly in handling noisy, or missing data—underscoring its suitability for diverse and challenging operational environments. The number of false positives for the drone is 5, as shown in Figure 9, despite being tested on fewer samples (176 in total) compared to DWAF, due to a lack of data.

Table 7: Evaluation Metrics for the MMTF System

Case	Precision	Recall	F1-Score	Accuracy
Without Data Missing	96.56%	96.02%	96.03%	96.07%
With Data Missing	96.36%	96.01%	95.99%	96.02%

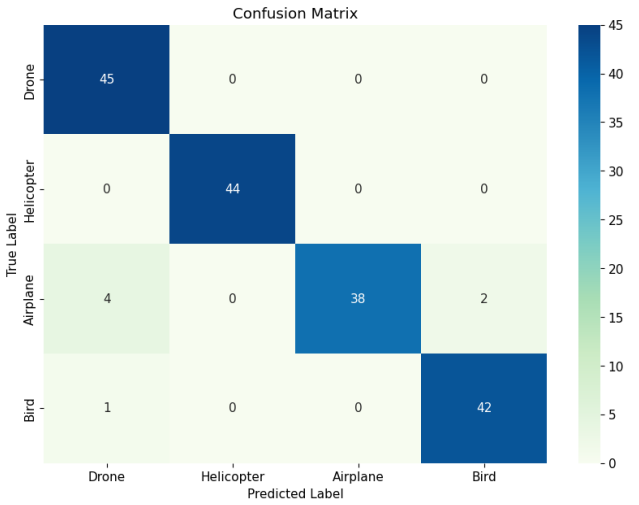

Figure 9: Confusion matrix of MMTF system showcasing performance with missing data.

Moreover, the comparative analysis of model accuracies reveals that the audio-only model achieved an accuracy of 81.95%, while the infrared (IR) video model registered 69.56%. The DWAF method significantly boosted accuracy to 89.31%, surpassing single-sensor approaches and highlighting its efficacy for real-world applications. The MMTF model, leveraging machine learning, outperformed on the test set of 176 samples, suggesting a promising avenue for future research, particularly with an expanded dataset.

In summary, while DWAF adeptly merges audio and video data, adjusting to varying performance levels and offering simplicity

Table 8: Comparison of proposed methods with existing methods.

Ref No.	Audio Data	Image Data	Distance (m)	No. classes	Accuracy (%)	F1-score (%)
[17]	✓	-	up to 60	2	62	-
[20]	✓	-	up to 4	1	70	-
[11]	✓	-	up to 150	2	-	80.09
[15]	✓	-	-	1	83	-
[30]	✓	-	-	5	81.3	-
[10]	-	✓	up to 25	2	-	96
[26]	-	✓	close	1	-	68.1
[31]	-	✓	-	2	-	79
[40]	-	✓	-	2	82.7	-
[28]	-	✓	-	2	-	74.2
DWAF approach	✓	✓	**up to 200**	**4**	**89.31**	**89.29**
[33]	✓	✓	up 200	4	78	-
[13]	✓	✓	up to 60	2	88.33	-
[14]	✓	✓	up to 100	1	92.53	-
MMTF approach	✓	✓	**up to 200**	**4**	**96.02**	**95.99**

and adaptability, it falls short in managing complex inter-modal interactions. In contrast, MMTF excels in integrating multimodal data through decision trees, adeptly navigating intricate patterns across modalities with superior accuracy and robustness, albeit at a little bit greater computational cost than DWAF.

5.4.3 Time Detection Performance. Our detection framework employs the MMTF method to efficiently process audio and video inputs, providing predictions for each segment in just 0.382 seconds per segment, while achieving a high accuracy rate of 96.02%. This framework is optimized for GPU processing, enhancing its suitability for real-time applications. In comparison, the alternative DWAF technique requires slightly less time for processing, at 0.356 seconds per segment.

5.4.4 Comparison with existing drone detection methods. A comprehensive comparison of all methods directly is challenging due to their evaluation on different datasets. However, one exception is the study identified in [33], which utilized the same dataset for evaluation. As illustrated in Table 8, we present a comparison between our proposed drone detection method and existing approaches. For consistency in evaluation, we adopted a similar k-fold validation approach as outlined in recent literature, setting k to 5. Our method demonstrated outstanding performance, achieving an accuracy of nearly 96.02% in detecting drones. The technique we introduced specifically addresses several challenges not considered in previous studies to the best of our knowledge. These include issues related to low resolution, various illuminations, occlusion, noisy audio, and missing data, which significantly impact detection accuracy.

6 CONCLUSION AND FUTURE WORK

Our research introduces a system that integrates audio and IR imaging data through an MMTF approach, surpassing the DWAF approach in performance. This system synergizes the strengths of a CRNN for audio processing with the YOLOv5s for IR image analysis, demonstrating enhanced efficiency in drone detection. Notably, the CRNN exhibits significant improvements over conventional LSTM models in audio recognition. Concurrently, YOLOv5s strike an optimal balance between speed and accuracy, outperforming both YOLOv3-tiny and the standard YOLOv5 in visual detection.

Our study's results confirm the efficacy of integrating multimodal data for drone detection, underscoring the transformative impact of machine learning algorithms on security and surveillance applications. Future directions for enhancing the system's adaptability and precision in various operational environments include integrating additional sensory modalities such as radar and LiDAR and testing them on a more extensive dataset to further refine the system's detection capabilities. Moreover, an interesting avenue for future research involves the exploration of cooperative distributed detectors. These detectors, equipped with diverse data sources, would communicate with each other to feed an intelligent centralized node. This setup could significantly improve detection capacity at greater distances and in challenging conditions, such as in the presence of potential obstructions or obfuscation attacks aimed at misleading detection efforts. This development aims not only to enhance the system's adaptability and precision but also to set the stage for more advanced and dependable drone detection technologies.

7 ACKNOWLEDGEMENT

This research was made possible with support from the Horizon Europe research and innovation programme of the European Union, under grant agreement number 101092912 (project MLSysOps).

REFERENCES

[1] 2023. FreeSound. https://freesound.org Online sound library.
[2] Mohammad F. Al-Sa'd, Abdulla Al-Ali, Amr Mohamed, Tamer Khattab, and Aiman Erbad. 2019. RF-based drone detection and identification using deep learning approaches: An initiative towards a large open source drone database. *Future Generation Computer Systems* 100 (2019), 86–97.
[3] Mohammed Aledhari, Rehma Razzak, Reza M. Parizi, and Gautam Srivastava. 2021. Sensor Fusion for Drone Detection. In *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. 1–7. https://doi.org/10.1109/VTC2021-Spring51267.2021.9448699
[4] Burchan Aydin and Subroto Singha. 2023. Drone Detection Using YOLOv5. *Eng* 4, 1 (2023), 416–433.

[5] Joël Busset, Florian Perrodin, Peter Wellig, Beat Ott, Kurt Heutschi, Torben Rühl, and Thomas Nussbaumer. 2015. Detection and tracking of drones using advanced acoustic cameras. In *Unmanned/Unattended Sensors and Sensor Networks XI; and Advanced Free-Space Optical Communication Techniques and Applications*, Edward M. Carapezza, Panos G. Datskos, Christos Tsamis, Leslie Laycock, and Henry J. White (Eds.), Vol. 9647. International Society for Optics and Photonics, SPIE, 96470F.

[6] Pietro Casabianca and Yu Zhang. 2021. Acoustic-Based UAV Detection Using Late Fusion of Deep Neural Networks. *Drones* 5, 3 (2021).

[7] Qiushi Dong, Yu Liu, and Xiaolin Liu. 2023. Drone sound detection system based on feature result-level fusion using deep learning. *Multimedia Tools and Applications* 82, 1 (Jan. 2023), 149–171.

[8] L. C. Danielle Furfaro and N. Musumeci. 2017. Civilian drone crashes into army helicopter. (2017). https://tinyurl.com/tyv3ayak

[9] Ungati Ganapathi and M. Sabarimalai Manikandan. 2020. Convolutional Neural Network Based Sound Recognition Methods for Detecting Presence of Amateur Drones in Unauthorized Zones. In *Machine Learning, Image Processing, Network Security and Data Sciences*, Arup Bhattacharjee, Samir Kr. Borgohain, Badal Soni, Gyanendra Verma, and Xiao-Zhi Gao (Eds.). Springer Singapore, Singapore, 229–244.

[10] Fatih Gökçe, Göktürk Üçoluk, Erol Şahin, and Sinan Kalkan. 2015. Vision-Based Detection and Distance Estimation of Micro Unmanned Aerial Vehicles. *Sensors* 15, 9 (2015), 23805–23846. https://doi.org/10.3390/s150923805

[11] Sungho Jeon, Jong-Woo Shin, Young-Jun Lee, Woong-Hee Kim, YoungHyoun Kwon, and Hae-Yong Yang. 2017. Empirical study of drone sound detection in real-life environment with deep neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)*. 1858–1862. https://doi.org/10.23919/EUSIPCO.2017.8081531

[12] Honggu Kang, Jingon Joung, Jinyoung Kim, Joonhyuk Kang, and Yong Soo Cho. 2020. Protect Your Sky: A Survey of Counter Unmanned Aerial Vehicle Systems. *IEEE Access* 8 (2020), 168671–168710. https://doi.org/10.1109/ACCESS.2020.3023473

[13] Juann Kim, Youngseo Kim, Heeyeon Shin, Yaqin Wang, and Eric T Matson. 2023. How Far Can a Drone be Detected? A Drone-to-Drone Detection System Using Sensor Fusion.. In *ICAART (3)*. 877–884.

[14] Juann Kim, Dongwhan Lee, Youngseo Kim, Heeyeon Shin, Yeeun Heo, Yaqin Wang, and Eric T. Matson. 2022. Deep Learning Based Malicious Drone Detection Using Acoustic and Image Data. In *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*. 91–92.

[15] Juhyun Kim, Cheonbok Park, Jinwoo Ahn, Youlim Ko, Junghyun Park, and John C Gallagher. 2017. Real-time UAV sound detection and analysis system. In *2017 IEEE Sensors Applications Symposium (SAS)*. IEEE, 1–5.

[16] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. 2022. CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer. arXiv:2209.06535 [cs.CV]

[17] Sayan Mandal, Lei Chen, Vishwa Alaparthy, and Mary L Cummings. 2020. Acoustic detection of drones through real-time audio attribute prediction. In *AIAA Scitech 2020 Forum*. 0491.

[18] MathWorks. 2023. Get Started with the Video Labeler - MATLAB & Simulink. https://fr.mathworks.com/help/vision/ug/get-started-with-the-video-labeler.html.

[19] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *SciPy*. https://api.semanticscholar.org/CorpusID:33504

[20] József Mezei and András Molnár. 2016. Drone sound detection by correlation. In *2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 509–518.

[21] Thomas Müller, Heiko Widak, Matthias Kollmann, Aleksej Buller, Lars Wilko Sommer, Raphael Spraul, Alexander Kröker, Ilja Kaufmann, Angelika Zube, Florian Segor, Thomas Perschke, Alina Lindner, and Igor Tchouchenkov. 2022. Drone detection, recognition, and assistance system for counter-UAV with VIS, radar, and radio sensors. In *Automatic Target Recognition XXXII*, Riad I. Hammoud, Timothy L. Overman, Abhijit Mahalanobis, and Kristen Jaskie (Eds.), Vol. 12096. International Society for Optics and Photonics, SPIE, 120960A.

[22] Zohaib Mushtaq and Shun-Feng Su. 2020. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics* 167 (2020), 107389.

[23] Phuc Nguyen, Hoang Truong, Mahesh Ravindranathan, Anh Nguyen, Richard Han, and Tam Vu. 2017. Matthan: Drone presence detection by identifying physical signatures in the drone's rf communication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 211–224.

[24] Seongjoon Park, Hyeong Tae Kim, Sangmin Lee, Hyeontae Joo, and Hwangnam Kim. 2021. Survey on Anti-Drone Systems: Components, Designs, and Challenges. *IEEE Access* 9 (2021), 42635–42659. https://doi.org/10.1109/ACCESS.2021.3065926

[25] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. 2019. Deep Learning for Audio Signal Processing. *IEEE Journal of Selected Topics in Signal Processing* 13, 2 (2019), 206–219. https://doi.org/10.1109/JSTSP.2019.2908700

[26] Chong Yu Quan, Ong Le Wei Edmond, and Sutthiphong Srigrarom. 2021. Identification of drone thermal signature by convolutional neural network. In *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*. 63–70. https://doi.org/10.1109/ICUAS51884.2021.9476825

[27] James Robert, Marc Webbie, et al. 2018. Pydub. http://pydub.com/

[28] Ulzhalgas Seidaliyeva, Daryn Akhmetov, Lyazzat Ilipbayeva, and Eric T Matson. 2020. Real-time and accurate drone detection in a video with a static background. *Sensors* 20, 14 (2020), 3856.

[29] Ulzhalgas Seidaliyeva, Lyazzat Ilipbayeva, Kyrmyzy Taissariyeva, Nurzhigit Smailov, and Eric T Matson. 2023. Advances and Challenges in Drone Detection and Classification Techniques: A State-of-the-Art Review. *Sensors* 24, 1 (2023), 125.

[30] Lin Shi, Ishtiaq Ahmad, YuJing He, and KyungHi Chang. 2018. Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments. *Journal of Communications and Networks* 20, 5 (2018), 509–518. https://doi.org/10.1109/JCN.2018.000075

[31] Subroto Singha and Burchan Aydin. 2021. Automated Drone Detection Using YOLOv4. *Drones* 5, 3 (2021). https://doi.org/10.3390/drones5030095

[32] Fredrik Svanström, Fernando Alonso-Fernandez, and Cristofer Englund. 2021. A dataset for multi-sensor drone detection. *Data in Brief* 39 (2021), 107521.

[33] Fredrik Svanström, Fernando Alonso-Fernandez, and Cristofer Englund. 2022. Drone Detection and Tracking in Real-Time by Fusion of Different Sensing Modalities. *Drones* 6, 11 (2022).

[34] Fredrik Svanström, Cristofer Englund, and Fernando Alonso-Fernandez. 2021. Real-Time Drone Detection and Tracking With Visible, Thermal and Acoustic Sensors. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 7265–7272.

[35] Vitalii Tyurin, Oleksii Martyniuk, Volodymyr Mirnenko, Pavlo Open'ko, and Ilona Korenivska. 2019. General Approach to Counter Unmanned Aerial Vehicles. In *2019 IEEE 5th International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD)*. 75–78. https://doi.org/10.1109/APUAVD47061.2019.8943859

[36] Ultralytics. 2020. YOLOv5. https://github.com/ultralytics/yolov5 GitHub repository.

[37] Ultralytics. 2021. YOLOv3: Real-Time Object Detection. https://github.com/ultralytics/yolov3. Accessed: September, 2023.

[38] Dana Utebayeva, Lyazzat Ilipbayeva, and Eric T. Matson. 2023. Practical Study of Recurrent Neural Networks for Efficient Real-Time Drone Sound Detection: A Review. *Drones* 7, 1 (2023).

[39] Jian Wang, Yongxin Liu, and Houbing Song. 2021. Counter-Unmanned Aircraft System(s) (C-UAS): State of the Art, Challenges, and Future Trends. *IEEE Aerospace and Electronic Systems Magazine* 36, 3 (2021), 4–29. https://doi.org/10.1109/MAES.2020.3015537

[40] Zizhe Wang, Lin Qi, Yun Tie, Yi Ding, and Yang Bai. 2018. Drone Detection Based on FD-HOG Descriptor. In *2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. 433–4333. https://doi.org/10.1109/CyberC.2018.00084

[41] Ghazlane Yasmine, Gmira Maha, and Medromi Hicham. 2022. Survey on current anti-drone systems: process, technologies, and algorithms. *International Journal of System of Systems Engineering* 12, 3 (2022), 235–270.