Deep Learning Models for Predicting Drone Sound Distances: Lightweight, Fusion and Hybridization Approaches

Dana Utebayeva *, Lyazzat Ilipbayeva *, Ulzhalgas Seidaliyeva *, Assel Yembergenova, Eric T. Matson

Posted Date: 28 October 2024

doi: 10.20944/preprints202410.2156.v1

Keywords: UAV sound distance; CNNs; RNNs; SimpleRNN; LSTM; BiLSTM; GRU; CNN-BiLSTM; UAV sound classification; Kapre method; melspectrogram; deep learning; drone sound detection; real-time UAV sound detection; fusion; voting system; hybrid models

 doi:10.20944/preprints202410.2156.v1

Article

Deep Learning Models for Predicting Drone Sound Distances: Lightweight, Fusion and Hybridization Approaches

Dana Utebayeva [1,*], Lyazzat Ilipbayeva [2], Ulzhalgas Seidaliyeva [1], Assel Yembergenova [3] and Eric Matson [4]

[1] Department of ET and ST, Satbayev University, Almaty 050013, Kazakhstan;
[2] Department of RET, International IT University, Almaty 050040, Kazakhstan;
[3] Department of CS, SDU University, Kaskelen, Kazakhstan;
[4] Department of CIT, Purdue University, West Lafayette, USA;
* Correspondence: d.utebayeva@satbayev.university

Abstract: In recent years, the widespread use of drones in daily life and large public events has raised serious safety concerns, especially due to incidents, both intentional and accidental. One of the most important aspects to prevent these risks is the ability to detect and accurately predict the distance of UAVs (Unmanned Aerial Vehicles) from people and restricted areas. One of the pressing issues is to develop systems to monitor their flights in restricted areas and predict suspicious movements in cases where suspicious drone flights are detected, which may be launched for video reconnaissance and information theft purposes. The advancement of acoustic-based recognition systems is growing with the development of deep learning. This study explores deep learning model architectures for the task of predicting UAV distances based on their flight sounds. The objective of this study is to determine whether sound-based classification can effectively predict drone movements at different distances from acoustic sensor points as they move from one area to another. Our experimental tests tried to predict the movement of UAVs based on the classification into 3 main zones. The results showed that drone sounds could be reliably detected in the movements between zones with an average recognition accuracy of 90 % using the hybrid CNN-BiLSTM model. Moreover, the implementation of such advanced acoustic sensor systems for UAV detection can improve the accuracy of real-time prediction, especially when integrated into a multimodal system with multi-sensor fusion methods.

Keywords: UAV sound distance; CNNs; RNNs; SimpleRNN; LSTM; BiLSTM; GRU; CNN-BiLSTM; UAV sound classification; Kapre method; melspectrogram; deep learning; drone sound detection; real-time UAV sound detection; fusion; voting system; hybrid models

1. Introduction

The use of drones has increased significantly in a number of industries as a result of technological advances, but this has also increased the dangers and accidents associated with their use. To address these issues, researchers around the world are working hard to create strategies and solutions that will improve the safety and regulation of drone use [1].

The range of applications of drones in everyday life has increased due to the rapid development of modern technologies and the improvement of their technical capabilities. However, drones are often used for both illegal and peaceful purposes without permission. The use of drones for reconnaissance and the collection of confidential information has observed in recent years, posing a great threat to both national and international security. Numerous examples have demonstrated the need to pay attention to this study area. For instance, the Canadian women's soccer team was accused of using drones to unlawfully record New Zealand training sessions during the 2024 Summer Olympics in Paris. The New Zealand Olympic Committee formally complained about the incident, which caused international controversy. The incident made clear how dangerous it is to use drones to spy on rival teams during athletic events [2,3]. In addition, the SpaceX center in Texas was recorded with another similar threat from a man. The county sheriff's department filed the incident after receiving a tip

 doi:10.20944/preprints202410.2156.v1

from SpaceX security. The drone owner said he was trying to see rockets near the launch site and said he didn't know it was a restricted area. The man was arrested for attempting to fly a drone at a strategically important facility. This incident highlights how important the drone threat is when it comes to protecting critical infrastructure [4]. A number of illegal drone flights in the White House area in the US have also hampered security measures [5]. So, drones equipped with high-quality cameras can also be used to remotely eavesdrop on individuals, companies, and government organizations. This concern may be related to the dangers of drones, such as spying or invasion of privacy, despite privacy claims [6]. These mentioned incidents have demonstrated the importance of monitoring drones and taking timely action against unauthorized flights over strategically important objects, especially their range over these buildings.

Hence, the recent examples listed above not only demonstrate that drones can pose a risk to vital facilities and infrastructures, but also demonstrate that they can be used for purposes other than visual monitoring, such as stealing confidential data or causing damage. In this sense, real-time intelligent systems that track drone movements in the air of a protected area and quickly identify any unauthorized use can highlight the importance of a security alert plan. It is essential to implement specialized drone control systems and set up systems that can predict their approach to restricted areas to ensure such security measures. That is, their distances or suspicious flights in the area must be determined in a timely manner. Effective and inexpensive remote control systems can significantly reduce the risk and stop illegal drone flights without large information losses. Therefore, estimating the distance between drones and specific locations or predicting the distance of UAVs in certain zones is one of the important goals of this work. Recent developments in deep learning-based audio classification have demonstrated that acoustic intelligent systems can predict the distance of UAV sounds. These advances demonstrate the possibility of incorporating intelligent UAV detection systems into the security of protected areas, especially in high-risk areas where spatial awareness is essential.

So, the study aims to explore deep learning models for drone distance recognition and prediction. In other words, the viability of distance recognition using drone audio signals will be investigated by experimentally evaluating different structures and comparing their characteristics. In this regard, the following objectives were set:

1) Analysis and exploration of deep learning models for distance prediction of drone audio signals.

2) Experimental review of different deep learning architectures using RNN and CNN neural networks, as well as hydrid models and weighed voting system. 3) Selection of the most effective architecture among the models and evaluation of its performance.

4) Investigation of different approaches to improve the accuracy and efficiency of distance estimation based on drone audio features.

5) Evaluation of the advantages and limitations of different drone audio recognition system architectures.

2. Related Works

Deep learning methods have shown their effectiveness in recognizing and classifying drone audio signals in studies [1,6–9]. In particular, CNN, RNN, and hybrid models of these two networks are among the neural network architectures that have been widely studied and used in this field in recent years. These models have demonstrated good performance in the tasks of predicting various drone flight states and their binary or multi-class classification. The tasks of predicting distances and improving their classification using drone signals are one of the emerging areas. Improving multi-layer architectures and structures optimized for sequential processing is also a priority area.

2.1. General Acoustic Sensor (AS) Systems

The authors Nijim et al. [10] primarily focused on identifying and categorizing drones by studying their sounds to address safety risks. Their method recorded drone sounds using acoustic sensors and

 doi:10.20944/preprints202410.2156.v1

classified the data using hidden Markov models (HMMs). They examined the sound patterns of drones such as the DJI P3 and Quadcopter FPV 250 during different flight phases (e.g., hovering, flying). The work used techniques such as data mining and clustering. The two main tools for detecting distinctive sound patterns were spectrograms and frequency analysis. Despite the potential of the method, the study highlights the need for a larger sound database, real-time processing, and improved efficiency for broader drone identification.

The work [11] presented the Drone Acoustic Detection System (DADS) from Stevens Institute of Technology. The system used propeller noise to detect, track, and classify UAVs. Using multiple microphone arrays, the system determined the direction of arrival (DOA) and localizes drones using triangulation. DADS consisted of microphone nodes in a tetrahedral configuration, processing audio data in real time. The system has been tested on drones such as the DJI Phantom 4, M600, Intel Falcon 8+, and DJI S1000. Limitations of the system included shorter range compared to radar and RF systems, and dependence on ambient noise. The maximum detection range reached 300 meters in a quiet environment.

The authors Cheranyov et al. in their work [12] considered ways to detect drones by acoustic method along with other methods. The passive acoustic method is based on the recognition of the sounds produced by the engines and structural elements of drones. This method relies on the sound library of known drones and compares the sounds coming into the system with them. Microphone arrays are used to monitor the trajectory of the drone, by this method the sound propagation angle is calculated and the direction of the radiation source is determined. Disadvantages of the proposed method in this work are small detection distances in noisy environments. It was concluded that powerful computers are needed for processing high-speed signals. In addition, an acoustically active method is also provided. It uses ultrasound waves to analyze signals reflected from drones using the Doppler effect. This method is distinguished by effective reflection from plastic parts and the ability to detect low-altitude objects, but its distance does not exceed 15 meters. The overall work suggested that the method of combining acoustic, passive radar and optical methods for drone recognition is promising.

The authors in [13] considered the integration of low-cost sensors for unmanned aerial systems (UAS) control and landing procedures. The work investigated the implementation of accurate distance calculation and detection of obstacles by integrating sensors for small quadcopters. Sonic Ranging Sensor (SRS) and InfraRed Sensor (IRS) were used in the study. SRS determines the distance using ultrasonic waves, but there can be errors due to the influence of air temperature and humidity. IRS is an infrared sensor that works with high accuracy, but signal nonlinearity can cause problems. In general, the sensors were controlled by an Arduino Mega 2560 microcontroller. These sensors collected data every 0.5 seconds and send them to the UAS control system. In the work, the authors used data integration algorithms to determine the approach of the aircraft to obstacles and ensure safety during landing. The measurement results of the sensors were combined and the distance is determined with the minimum deviation. However, this work only focused on the development of a reliable range and direction detection system for small quadcopters. In the future, authors planned to increase the accuracy of the system by increasing the number of sensors and introducing complex algorithms.

The study [14] proposed an approach based on Euclidean distance to detect acoustic signals of drones (UAVs). In the work, Short Time Fourier Transform (STFT) was used to analyze the time-dependent frequency composition of sound. Euclidean distance (ED) was used to distinguish drone sounds from other sounds by comparing acoustic characteristics. This method can be used in the field of security to monitor drones and detect illegal flights. In general, the proposed method was aimed at effective recognition of acoustic signals of UAVs, which may be especially important for security applications. However, the drone's recognition area was not specified. And the authors concluded that it was possible to achieve a reliable system by increasing the database.

In the work [15], a system for detecting and analyzing unmanned aerial vehicles (drones) by sound was presented. The main goal of the proposed system was real-time detection and tracking of

 doi:10.20944/preprints202410.2156.v1

drones using low-cost microphones and audio data. The system processed audio signals using the Fast Fourier Transform (FFT) method and performed detection using two different algorithms: Plotted Image-Based Machine Learning (PIL) and k-Nearest Neighbors (KNN). The system was distinguished by its real-time operation, the ability to continuously recorded sounds through a microphone and convert them into frequencies using FFT, used inexpensive devices, and used sound-based detection as an alternative to high-precision cameras. PIL method: calculated structural similarity indices using images from the FFT plot. This method achieved 83 % accuracy. KNN method Converts audio data into CSV files and analyzes it by nearest neighbor detection. The accuracy of this method is 61 %. In the work, the use of additional algorithms (for example, Convolutional Neural Networks) to increase the accuracy of detection was planned as a future work. It was intended to improve the overall system by introducing ways to increase data and remove noise. This system tried to provide a simple, cheap and effective alternative way to detect drones. The area of drone detection was not specified.

2.2. Acoustic Sensor Systems Based on Machine Learning and Deep learning

The work [16] examined the sound recognition technologies of quadcopter-type unmanned aerial vehicles (UAV). The authors proposed recognition of different types of UAVs using deep neural networks (CNN), identifying features by binaural presentation of sounds using the Mel spectrum. The study compared UAV sounds with other types of environmental noise, such as bird calls and vehicle sounds. The authors conducted comparative analysis with various audio features such as Melspectrogram, MFCC and CNN, SVM models. The results of the study confirmed that the CNN model based on Melspectrogram performed better than other method. It is concluded that the proposed method allows accurate recognition of quadcopters based on sound and helps to effectively distinguish UAVs from ambient noise. It was said that the next steps of the research will be the recognition of UAV sounds in different flight states and the processing of long-range sounds.

And in works [7,8], Melspectrogram and RNN networks have achieved successful drone sound recognition accuracy. And in this study, the proposed system was introduced as an effective recognition method consisting of a lightweight architecture. In our previous work, which was in the direction of intelligent UAV distance estimation system [8], we tried to explore a deep learning model for determining the distance of unmanned aerial vehicles (UAVs) using their audio signals. The study was conducted on the recognition system of UAV sound at different heights (from 5 to 50 meters) using the GRU (Gated Recurrent Unit) neural network. During the research, sound signals were analyzed with 94 % accuracy at a distance of every 10 meters, and with 98 % accuracy at every 15 meters. Drone flights have not been studied in less limited database applications and in full dynamic mode. Despite the limitations, the study has potential for future integration with multimodal sensor systems. The importance of this research is in the application of the method of processing sound signals through deep learning and in the formulation of the possibility of this method to recognize UAVs at a real distance.

In other studies [17–24] tried to investigate UAV sound recognition based on machine learning methods, and the works [1,9,25–38], attempts were made to use convolutional neural networks (CNNs) and recurrent neural networks (RNNs) during the study of various applications for drone sound recognition. Different hyperparameters of neural networks were used in these works. The results of their study showed that deep learning methods can be used to recognize drone sounds. The studies [6,39] have conducted an extensive and systematic review of the field of drone recognition using deep learning methods.

2.3. Weighted Voting Systems

In the study work [40], the use of ensemble classifiers was considered as a new approach to pattern recognition. The authors evaluated the effectiveness of the weighted pluralistic voting strategy and implemented it in face and voice recognition tasks. According to the authors, by properly balancing the elections, pluralistic voting can be used to improve the efficiency of various classifiers. The results of

 doi:10.20944/preprints202410.2156.v1

the study demonstrated that the use of ensemble classifiers based on weighted voting allows achieving high recognition rates.

The study [41] investigated at a reliable and simple-to-implement method for real-time voice activity detection (VAD). The proposed method took into consideration the a weighted voting system. Experiments demonstrated that the suggested method outperforms the baseline method in a number of areas, such as threshold selection and channel distortion. The authors claimed that using the proposed weighted voting approach, the average performance of VAD increased to 89.29 % for 5 different noise types and 8 SNR levels. The overall performance was 13.79 % higher than the SPVD-only approach and even 2.25 % higher than the non-weighted voting scheme.

The study [42] proposed a neural network (NN)-based weighted voting classification algorithm. The authors investigated that the performance of the algorithm was verified using real experimental data, and the results show that the proposed method has a higher accuracy in classifying the features of the target signal, achieving an average classification accuracy of about 85 % when using a deep neural network (DNN) and a deep belief network (DBN) as the base classifier. Their experiment showed that the NN-based weighted voting algorithm improved the accuracy of the target classification by about 5 % compared with a single NN-based classifier, but it also increases the memory and computation time required to run the algorithm.

Thus, the above-discussed studies have shown that various models of deep learning methods, especially CNN, RNN and their hybrids, play an important role in effectively building drone detection and sound classification systems. These systems significantly improve classification accuracy and provide robustness across a variety of scenarios, and weighted voting systems have also shown effectiveness in object recognition tasks. And the next section discusses the data preparation and deep learning model architectures used in this study.

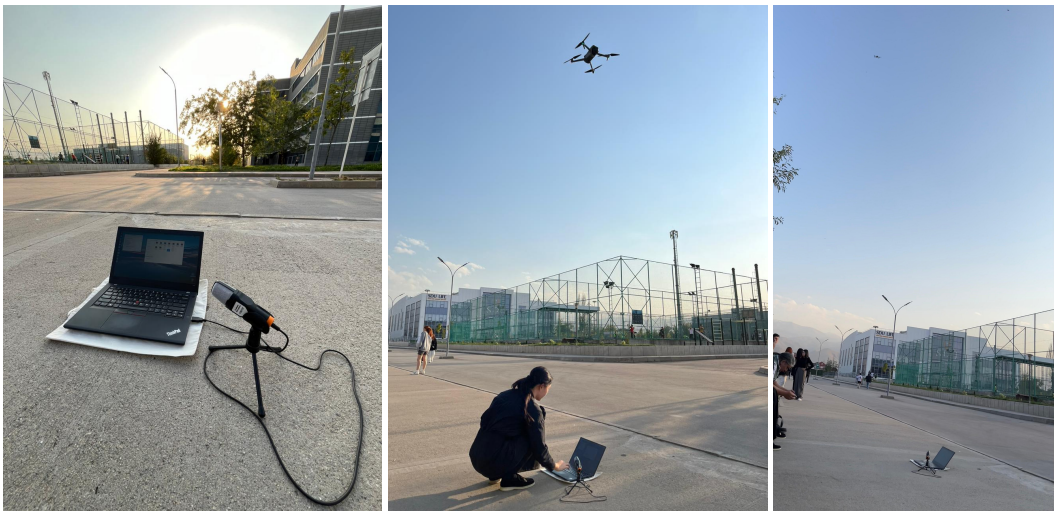
3. Proposed System

3.1. Data Preparation

Deep neural networks based on Recurrent and Convolutional neural networks play an important role in the field of audio data processing due to their ability to recognize complex patterns in huge amounts of data as mentioned above. Data preparation is therefore one of the most important aspects of deep learning, as the ability of the model to learn effectively and provide accurate predictions is directly dependent on the state of the input data. This step is especially important when working with audio data, as signal processing and taking into account numerous external elements, such as background noise and other motorized objects, significantly affect the recognition capabilities of the neural model. This section focused on the preparation process of audio data recorded from unmanned aerial vehicles (UAVs) that were captured at different distances from the Acoustic Sensor Point (ASP), 1. UAV audio data is a complex type of data, since its acoustic characteristics can vary depending on many factors such as model, distance, speed, and environmental conditions. In this regard, careful data preparation becomes a critical prerequisite for the effective application of deep learning methods.

So, to develop the system of this work, it was tried to collect the data carefully and purposefully, representing the sounds of unmanned aerial vehicles (UAVs) recorded at different distances from the protected object, Figure 1. Secondly, since an important source of information for training the deep learning model are these audio recordings that capture the acoustic signatures of drones at different distances. To adapt to different flight conditions, three different UAV models were used - DJI Mini 2, Qazdrone and DJI Air 3, in Table 1.

 doi:10.20944/preprints202410.2156.v1



(**a**) (**b**) (**c**)

Figure 1. UAV distance sound data preparation: a) Microphone placement; b) UAV flight in close zone; c) UAV flight in far zone.

Several factors led to the selection of these models for the study. First, they are often used for video recording. Since these three models were available at the time of the experiment, the choice was intentionally limited to only three models, Table 1. It should be noted that the current study did not use the existing database containing drone sound recordings from our earlier study [7]. This choice was made for a number of reasons. First of all, the data collected in the earlier study [7] did not meet the requirements of this assignment. It should also be noted that the data selection in the previous study was driven by its specific research question, which was intended to address a different problem. The lack of detailed information on the distances at which records had previously been made was another major drawback of the previous database for this work's system. Thus, for supervised learning in this work, specific distance specifications are needed, but these were missing in earlier records. Therefore, to successfully complete the work outlined in the current study, it was necessary to create a new database adapted to investigate the acoustic properties of UAVs from specific distances.

Table 1. Technical parameters of used UAV models.

UAV models	Parameters	Flying Distance (in meters)
DJI Mini 2	Mode "S": 3.5-5 m/s Mode "N": 3 m/s Mode "C": 1.5-2 m/s	1–50
DJI Air 3	Max ascent and descent speed: 10 m/s Max Horizontal Speed (at sea level, no wind): 21 m/s	2–50
Qazdrone		1–50

Our models were developed to process audio data in the "WAV" format. All recordings of unmanned aerial vehicles (UAVs) were made with a microphone with a resolution of 16 bits and a sampling rate of 44,100 Hz (22,050 in some recordings). The UAVs were recorded in various flight modes, including reciprocating, vertical (up and down) movements and at different speeds, starting from positions in close proximity to the microphone, but at specified distances or between specified distances in motion, Tables 2 and 3.

To record audio data from unmanned aerial vehicles (UAVs), two different recording methods were used, each of which assumed unique flight conditions and drone motion modes, which ensured a variety of UAV audio recording scenarios. The first method assumed recording sounds in a semi-dynamic mode, in which the drones were in a state of relatively limited mobility in a vertical position,

 doi:10.20944/preprints202410.2156.v1

but at the same time, flight variability was preserved. That is, with this method, the UAVs flew at a certain altitude with minimal deviations, but in motion along the horizontal axis at a diagonal of 2-3 meters.

Table 2. Parameters of sound recordings during dynamic movement of "DJI Mini 2" drone.

Zones	distances from the ASP	Speed	Actions
Zone 1	2-5 meters	4-5 m/s	up and down; straight; circling
	6-10 meters	7-8 m/s	up and down; straight; circling
	11-15 meters	3-5 m/s	up and down; straight; circling
Zone 2	16-20 meters	9-10 m/s	up and down; straight; circling
	21-25 meters	9-10 m/s	up and down; straight; circling
	26-30 meters	8-10 m/s	up and down; straight; circling, loud noises
Zone 3	31-35 meters	9-10 m/s	up and down; straight; circling
	36-40 meters	3-4 m/s	up and down; straight; circling
	41-45 meters	4-5 m/s	up and down; straight; circling (football players screaming was parallelly)
	46-50 meters	4-5 m/s	up and down; straight; circle; (football players noise)

Here, the noise of repair work and the active conversation of a large group of students were very close. The mini-truck unloaded at a very close area.

The second recording method assumed a full dynamic flight model, in which the drones moved at different speeds and in different directions, simulating real drone operating conditions in different scenarios. In particular, in this mode, the drones made up and down movements, which ensured the recording of sounds at different altitudes, but with the preservation of certain micro-altitudes (for example, from one to five meters).

Table 3. Parameters of sound recordings during dynamic movement of "DJI Air 3" drone.

Zones	distances from the ASP	Speed	Actions
Zone 1	2-5 meters	4-5 m/s	up and down; straight; circling
	6-10 meters	4-5 m/s	up and down; straight; circling
	11-15 meters	4-5 m/s	up and down; straight; circling
Zone 2	16-20 meters	2-10 m/s	up and down; straight; circling; back and forth
	21-25 meters	2-10 m/s	up and down; straight; circling; back and forth
	26-30 meters	2-10 m/s	up and down; straight; circling; back and forth
Zone 3	31-35 meters	3.5-15 m/s	up and down; straight; circling
	36-40 meters	3.5-15 m/s	up and down; straight; circling
	41-45 meters	3.5-15 m/s	up and down; straight; circling
	46-50 meters	3.5-15 m/s	up and down; straight; circle;

This approach allowed us to detail the acoustic data, since the sounds produced by drones in a static state and in motion can differ significantly. The frequency characteristics of sounds change depending on the dynamics of movement, since different vibration and noise effects are created when flying at different speeds and at different altitudes. Given this important feature, we intentionally recorded sounds in static, semi-dynamic and dynamic modes to account for the full range of possible acoustic variations characteristic of UAV flight. Recording audio data in this way provided deep learning models with maximum accuracy in recognizing and classifying sounds.

To better analyze and simulate the actual operating conditions of the system, UAV flights were planned under conditions divided into spatial zones such as Zone 1, Zone 2 and Zone 3. Zone 1 was

 doi:10.20944/preprints202410.2156.v1

taken as the closest zone to the protected point up to 15 meters. Zone 2 was assumed to be the middle zone to the guarded point from 16 meters to 30 meters. A little further, Zone 3 was adopted from a range of 31 meters to 50 meters, Figure 2.

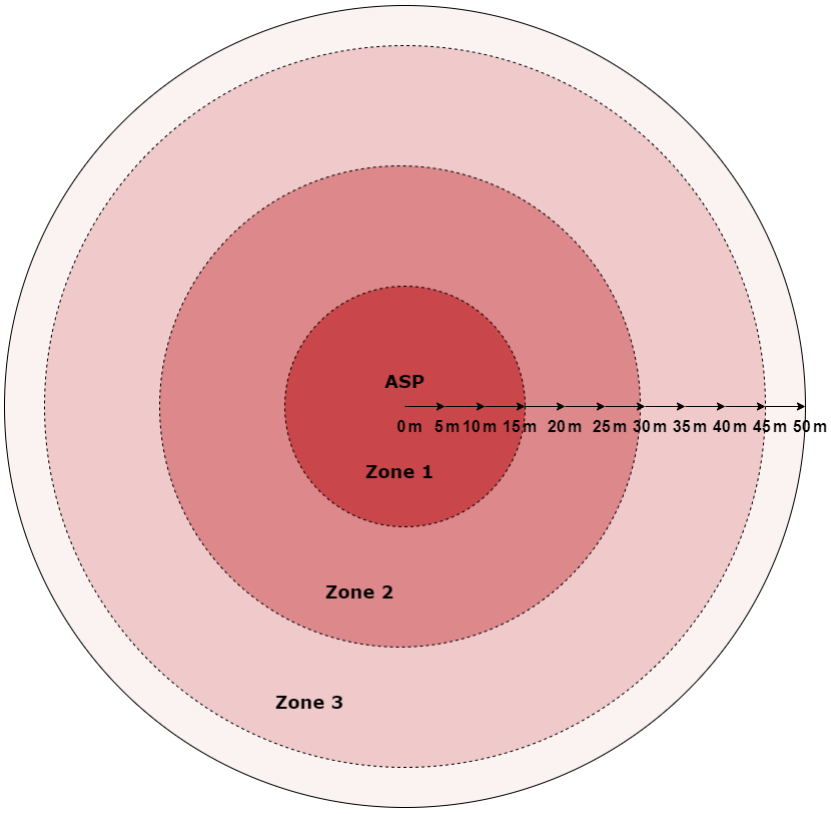


Figure 2. Simulation of zone-based acoustic sensor system for UAV distance based recognition system.

These three main flight zones were divided into micro-zones during the recordings with a one-meter step between the zones, which may allow for a more detailed recording of sounds at different altitudes in the case of future studies.

This structure of the zones contributed to a more accurate division of the sound data by altitude ranges, which in turn may allow for differences in acoustic characteristics that may occur at different flight levels. For example, the first micro-zone was recorded at altitudes from one to five meters, which included both low and medium drone flight levels. The second micro-zone was recorded at altitudes from six to ten meters, which allowed for the analysis of sounds at higher levels. Such a detailed division of zones with a one-meter step may allow the researchers to obtain more accurate and diverse data for subsequent analysis and for further studies on training deep learning models.

Thus, both fully dynamic and semi-dynamic flight mode sound recording methods captured a wide range of UAV-specific sound signatures under various flight conditions.

In addition, various engine-based sounds were specially recorded during the recording and added to the audio database as a separate "No UAV" class. In particular, the sounds of moving vehicles such as cars, compact trucks and motorcycles were also recorded. In addition, sounds from numerous construction and repair works were included, as well as loud ones, since loud construction noises were heard simultaneously during the UAV flights. Train sounds were also added to this class. The sounds of people moving and talking during a football match were also recorded to demonstrate the impact of the surrounding social and sporting events that were taking place during the recording. Sounds such as wind, rustling leaves and birds chirping were also added.

Thus, the "No UAV" class of the collected audio data includes various noises, including traffic noise of vehicles, social noises of public events, sounds of nature and maintenance noises. This data

 doi:10.20944/preprints202410.2156.v1

will be used in the future to accurately distinguish background and target sound and conduct effective analysis. This comprehensive approach allows for more accurate classification of various noises and their systematic processing.

All sound databases, that is, the names of classes based on zones and durations of sounds that were obtained from the three specified models, are fully displayed in the Table 4, and the "Dataavailability" section below offers access to these sounds.

Table 4. Composition of UAV sound dataset.

Name of Classes	Total Duration, (s)	Duration of Training Set, (s) Train Set 80 % Test Set 20 %	Duration of Validation Set, (s) (unseen data)
	16962	**15266**	**1696**
No UAV	3780	3402	378
UAV in Zone 1	3766	3390	376
UAV in Zone 2	3993	3593	400
UAV in Zone 3	5423	4881	542

3.2. Model Preparation

So, the main goal of this work is to study the possibility of recognizing and classifying sounds of unmanned aerial vehicles located at different distances using different deep neural network architectures. That is, special attention is paid to the possibility of recognizing UAV sounds to predict their distance from the protected point. This area of research is of particular importance, since successful classification of sounds at different distances can be applied in a number of practical scenarios, such as security monitoring or detection of unauthorized drones in specially protected areas with repetition of sensor nodes. To achieve the stated goal, several objectives were set in this work: studying various deep learning architectures of *single-layer*, *stacked*, and *hybrid models*.

The results of previous studies [7] have demonstrated the effectiveness of RNN (recurrent neural networks) and CNN (convolutional neural networks) architectures (Table 5) in recognizing UAV sounds. And in our previous study [7], RNN networks gave good results in some tasks. And in our other study, we noticed the optimality of hybrid RNN - CNN models: in particular, these models were able to successfully classify different UAV sound categories and their sounds in different states. In this regard, the main objective of the current work was to more deeply study the capabilities of RNN and CNN in drone sound recognition by exploring their different architectures and different combinations.

First of all, single-layer recurrent neural networks "1L" were investigated, which, despite their simplicity, can be very effective in some classification problems, Tables 6 and 7. Consequently, for a more in-depth analysis, Stacked architectures based on two-layer “2L” recurrent neural networks, which have a slightly larger number of layers, were also considered, Table 8.

A voting system based on a single-layer neural network was also one of the approaches that was explored in this work. This approach uses multiple models, each of which classifies separately, and the final result is chosen depending on how well these models agree. Therefore, in situations where the data contains noise or other distortions, the voting approach can reduce the probability of error and improve the overall classification accuracy.

The study further explored hybrid models, which learn overall classification performance by combining different networks, Tables 9 and 10. To maximize the benefits of each strategy, hybrid models often include components from several different neural networks, such as CNNs, RNNs, or networks of the same type. As a result, these models can better handle tasks involving highly variable audio signals and be more robust to changes in the data.

Overall, the study aimed to compare the performance of different architectures in different compositions and evaluated the advantages of their combinations in terms of the probability of recognizing UAV sounds at a distance. And the analysis of this research work based on a large number of experiments will help improve the accuracy of recognizing drone sounds at a distance, as well as

 doi:10.20944/preprints202410.2156.v1

identify effective models that can have a significant impact on future applications in this direction. In addition, the study explored the potential of using deep neural networks in various applications, such as real-time UAV range recognition for additional sensors in multimodal detection systems.

The study considered a number of experimental works built on different architectures. The hyperparameters of each of these model architectures are explained in the following subsections.

3.2.1. CNN Based Deep Learning Architectures

In this subsection, we discuss the creation of a Deep Learning model architecture using CNN neural networks. The task is performed based on several classifications of UAV acoustic distance data. And this UAV audio data is fed to the input of the model in the form of one-second audio files. Then the first layers process this audio data through a Melspectrogram layer in real time. The hyperparameters of Melspectrogram are given in Table 5. Then the normalization layer process and feed the data to the next 2D CNN layer.

Table 5. Optimization of hyperparameters of proposed Deep Learning models for the experimental work "CNN".

Layers	Parameter	Range
Melspectrogram	*Sampling rate*	*16,000 Hz*
	Window length	*512*
	Hop length	*160*
	Number of Mels	*128*
	(Frequency, Time)	128 * 100
LayerNormalization	Batch Normalization	
CNN 2D	cells	64
	kernel size	(3,3)
	activation	'tanh'
MaxPooling2D	pool size	(2,2)
	padding	'same'
CNN 2D	cells	128
	kernel size	(3,3)
	activation	'relu'
Flatten		
Dense	Dense	(# classes) 4
	Activation in classification	softmax
	Optimization solver	adam
	# epochs	18

Usually, deep neural networks apply a MaxPooling2D layer after a 2D CNN (Convolutional Neural Network) layer. And here, a MaxPooling2D layer was applied after a 2D CNN layer to extract meaningful features from the audio data. Usually, a CNN layer uses the data to create feature maps that can learn by displaying the meaningful information included in the processed data. This CNN and Maxpolling2D structure was repeated once more, and then a Flatten layer was added. In the last step, a Dense layer was added. The Dense layer was placed in the last step and was used to make final decisions based on the received data, i.e. to classify the data into certain classes or to predict the results. The selected hyperparameters of each layer of this architecture are fully presented in Table 5.

3.2.2. "1L" Single-Layer RNN-Based Lightweight Deep Learning Architectures

This category of our experiments tried to study models empirically using deep learning methods based on a single-layer (1L) recurrent neural network (RNN) architecture. The input layer of the model consisted of melspectrograms that processed UAV audio signals. That is, several RNN architectural types were investigated, such as SimpleRNN, LSTM, GRU, and BiLSTM. The Kapre library was used to process the data provided for the first layer. Overall, this architecture is based on our previous research, as the research we have conducted has proven that similar architectures produce good results.

 doi:10.20944/preprints202410.2156.v1

Thus, the Melspectrogram layer was first adjusted using the hyperparameters listed as in Table 6, and then the LayerNormalization layer was applied. This layer applies a systematic scaling of the data provided to later layers and normalizes the activation in each feature map to ensure continuous model training.

Table 6. Optimization of hyperparameters of proposed Deep Learning models for a series of experimental works "1L RNNs".

Layers	Parameter	Range
Melspectrogram	*Sampling rate* *Window length* *Hop length* *Number of Mels* (Frequency, Time)	*16,000 Hz* *512* *160* *128* 128 * 100
LayerNormalization	Batch Normalization	
Reshape	TimeDistributed (Reshape)	
Dense	TimeDistributed (Dense), tanh	128
SimpleRNN/LSTM BiLSTM/GRU	(cells)	128
concatenate	TimeDistributed (Dense)tanh; SimpleRNN/LSTM/BiLSTM/GRU	
Dense	Dense, ReLU	(64)
MaxPooling	MaxPooling1D	
Dense	Dense, ReLU	(32)
Flatten		
Dropout	Dropout	0.5
Dense	Dense, ReLU activity regularizer	32 0.000001
Dense	Dense Activation in classification Optimization solver # epochs	(# classes) 4 softmax adam 18

The Melspectrogram input has been transformed into the format needed for a recurrent neural network by adding a Reshape Layer (TimeDistributed Reshape). The TimeDistributed wrapper allows the reformatting operation to be performed in multiple time steps. Later, a 128-unit TimeDistributed Dense layer with “tanh” activation was applied. This process helps the RNN layer extract the desired features from the input data. One of the SimpleRNN, LSTM, GRU or BiLSTM layers was used as the primary recurrent component. These layers are responsible for processing the features of the audio signals of the UAV states. The results of the previous TimeDistributed Dense layer with “tanh” activation and the recurrent layers were combined using the Concatenate layer, which optimized the network structure. Here, we observed attempts to merge various layers. Nevertheless, a successful outcome was obtained by combining data from neural network layers with the TimeDistributed Dense layer and "tanh" activation layer. After the RNN layer, the model was passed through a dense layer. At this stage: a dense layer with 64 units was chosen based on a series of empirical experiments.

Further, a MaxPooling1D layer was used to improve the efficiency of the model and reduce the possibility of overfitting. This layer provides computational efficiency and reduces the size of the feature map. To feed the input data to the final classification layers, a Flatten layer converts the multidimensional data into a one-dimensional vector. A Dropout layer was used and 50% of the neurons were randomly turned off to avoid overfitting. The ability of the model to adapt to new, untested data is improved by this technique.

A pair of Dense layers were used as output; the final refinement is performed by a dense layer with 32 units. The probability for each of the four classes was predicted by adding a dense layer with a

 doi:10.20944/preprints202410.2156.v1

softmax activation function. Since the Adam optimizer works well with recurrent networks that have an adaptive learning rate, it was used to optimize the model parameters. To control overweights and prevent overfitting, a value of 0.000001 was chosen as the regularizer coefficient. 18 epochs were used for the training phase. The final softmax layer can determine the probability of each class and outputs the categorization results of the model. In our prior research, the model was able to reliably identify audio signals and detect changes in drone type or payload with the help of the proposed architecture. Meanwhile, it will be investigated in the current study for UAV distance sounds. However, since the database composition has changed this time, the actual hyperparameters have also changed. In addition, the Results section discusses the series of experiments.

In general, the research was carried out in two new ways while maintaining this architecture: one is to increase the number of cells in the RNN neural network while maintaining its structure, and the RNN network was used as two layers. The following table 7 illustrates how the architecture hyperparameters changed in the RNN cell augmentation architecture. However, as studies have shown, only LSTM, BiLSTM and GRU networks gave high enough recognition and prediction results, so only these two methods LSTM and GRU were created.

Table 7. Optimization of hyperparameters of proposed Deep Learning models for a series of experimental works "1L RNNs" with 256 cells.

Layers	Parameter	Range
Melspectrogram	...	...
...	...	...
GRU/LSTM	(cells)	256
...	...	...

LSTM and BiLSTM showed similar high recognition skills. LSTM was trained with less time, so only GRU and LSTM were tested to check the direction of RNN cell augmentation. Since evaluating these two approaches made the strategy useless, we attempted to save time by not testing other types of RNNs. And testing the RNN network as a two-layer network is explained in the section "2L stacked layer RNN-based Deep Learning architectures".

3.2.3. The Voting System

The prediction of UAV recognition by weighted voting using the fusion of various trained single-layer RNN models is another significant experimental study direction in this work. In a weighted voting system, participating models have different levels of influence or “weight” in the recognition system. Unlike traditional voting systems, where each participant typically receives an equal number of votes, weighted voting systems vary the importance of each vote based on predetermined weights. Assigning weights and combining model votes are the two main components of a weighted voting system. When assigning weights, each voter model is assigned a weight that reflects its influence. This weight can be determined by a variety of factors, such as the ability to recognize patterns in certain tasks. And each is weighted based on its reliability or performance. When making decisions, the votes of the models are summed, but instead of simply counting each vote equally, the votes are multiplied by their respective weights. The outcome is then determined by a weighted sum, with the option with the highest combined score usually winning.

In essence, the weighted voting system allows for the different degrees of influence of the participating models to be taken into account, making it a versatile and effective decision-making tool in situations where equality of votes does not accurately reflect the complexity of the problem.

In our experimental studies, the weights assigned to the participating models are explained in detail in the Results section. This choice was driven by the need to first evaluate how well RNN model type recognition skills performed on the UAV audio data we study. The Results section explains in detail the weights we used for our models and their impact.

 doi:10.20944/preprints202410.2156.v1

3.2.4. "2L" Stacked Layer RNN-Based Deep Learning Architectures

For further investigation, we constructed a 2L-layer stacked RNN-based structure by keeping the design shown in Table 6 and including two layers with an RNN network of 128 cells each, Table 8.

Table 8. Optimization of hyperparameters of proposed Deep Learning models for a series of experimental works "2L RNNs".

Layers	Parameter	Range
...	...	...
GRU/LSTM	(cells)	128
GRU/LSTM	(cells)	128
...	...	...

And the hybrid combination of RNN's own neural network types with each other or hybrid connection with CNN network is explained in the next section.

3.2.5. Hybrid Models

The final stage of the research focused in particular on hybrid models that combined types of recurrent neural network architectures with each other or with convolutional neural networks. Combining CNNs and RNNs creates hybrid models that combine the benefits of two different neural networks: while RNN analyzes temporal dependencies or the temporal order and long-term relationships in the data, CNN enables efficient spatial feature recognition or the processing of various structural patterns in the data. To open up new possibilities in UAV audio signal detection, this synergistic approach that combines CNN and RNN was explored.

First, as shown in Table 9, two different types of RNNs were tested in combination.

Table 9. Optimization of hyperparameters of proposed Deep Learning models for a series of experimental works "LSTM-GRU".

Layers	Parameter	Range
...	...	...
LSTM	(cells)	128
GRU	(cells)	128
...	...	...

Subsequent exploratory tests were conducted using a combination of CNN and RNN network types, Table 10.

 doi:10.20944/preprints202410.2156.v1

Table 10. Optimization of hyperparameters of proposed Deep Learning models for a series of experimental works "conv2D-LSTM" and "conv2D-LSTM".

Layers	Parameter	Range
Melspectrogram	...	...
LayerNormalization	Batch Normalization	
CNN 2D	cells kernel size activation	32 (3,3) 'relu'
MaxPooling2D	pool size	(2,2)
CNN 2D	cells kernel size activation	64 (3,3) 'relu'
MaxPooling2D	pool size	(2,2)
Reshape		
Dense	TimeDistributed (Dense), tanh	128
LSTM **(or BiLSTM stacked by GRU)** **or BiLSTM)**	cells (cells of each RNN) (cells)	128 (128) (128)
concatenate	TimeDistributed (Dense)tanh;	
Dense	Dense, ReLU	(64)
MaxPooling	MaxPooling1D	
Dense	Dense, ReLU	(32)
Flatten		
Dropout	Dropout	0.5
Dense	Dense, ReLU activity regularizer	32 0.000001
Dense	Dense Activation in classification Optimization solver # epochs	(# classes) 4 softmax adam 18

All directions of the research work were aimed at an in-depth analysis of various deep neural network architectures used in UAV sound recognition and expanding their capabilities. The main goal was to determine the most effective one by comparing the recognition advantages of different architectures. The use of all hybrid models was analyzed with experimental results compared with previous models. Through this study, the effectiveness of combinations of different architectures was evaluated, and it can be said that the foundation for future directions of UAV sound recognition systems was laid. This is because the Results section can clearly describe the effective model based on the obtained results. So, this work not only considered CNNs and RNNs with a hybrid connection structure, but also studied various types of RNN networks with a mutual hybrid structure among themselves. Their complete results are presented in the next section.

4. Experiment Results and Discussion

In this research work, experiments were conducted in five main directions. In the first stage, deep learning network architectures based on a single-layer recurrent neural network (RNN) were studied. Also, within the framework of the first direction, the CNN network was studied. Subsequently, the prediction performance of several relatively successful RNN-based models, namely GRU, LSTM, and BiLSTM, were evaluated using the voting method. They were examined using the trained forms in two and three different versions. In the third stage, the structure was studied by increasing the number of neuron cells in two different RNNs. These models were then considered with a two-layer stacking with neurons of the same size as in the first experiment. At the final stage, hybrid models of different neural

 doi:10.20944/preprints202410.2156.v1

networks were tested and a comprehensive analysis of their results was conducted. The comparison of results and conclusions of the analysis are presented in the "Discussion" section.

4.1. Training

A series of experiments were conducted using a simple laptop with an Intel(R) Core(TM) i5-8265U processor running at 1.60 GHz, using the Python programming language (Spyder integrated development environment). Kapre method libraries and other necessary layer libraries are preinstalled for processing and analyzing Melspectrograms and other architecture layers.

The models were built and trained at various epochs during the study's first phase, and each model's output from 25 to 50 epochs was thoroughly examined. The training results were carefully observed and the model performance was evaluated at each selected epoch number. By comparing the results, it was found that the models achieved a "good fit" to the training data after the 18th epoch, when the model performance peaked and then overfit. At epoch 18, it was found that the model parameters had stabilized and no further training was required.

After the training process was completed, all trained models were saved in special files with the .h5 extension. This format is suitable for subsequent easy loading of models and further use for predicting new data. Thus, model reuse processes not only make it possible to make predictions, but it was additionally convenient to apply these trained models to the voting method. Stored patterns can provide high efficiency in processing and predicting new audio data, which increases the practical relevance of research.

The original data was split up into three parts for the model training procedures. For the model training processes, the raw data was divided into three parts. The number of audio files of the original data was 16962 seconds or audio files. About 10 % of the audio files of each class in this database were saved separately in advance for validation after training. In general, each audio file length in the databases was adapted to a certain length of 1 second during data adaptation. Because the processing and recognition system is important every second for real time. And the remaining 90 %, that is, 15266 audio files, were divided into two parts as 80 % for training and the remaining 20 % for testing. This distribution of the database contributed to a comprehensive and objective assessment of the quality of the models, their effectiveness and ability to make predictions, as well as for the purpose of checking the performance of the system and monitoring the ability to work in real time. And the pre-prepared individual audio files, not included in the training and testing data, were used only during the validation period. Thus, the validation process was aimed at identifying possible system errors and assessing the reliability and stability of its results.

The hyperparameter selection of the melspectrogram layer for real-time audio data processing has been discussed in detail in our previous studies [7]. However, the main objective of this research work is to experimentally investigate deep learning architectures for processing UAV audio data from different distances in five different directions as mentioned above.

4.1.1. Deep Learning Models Based on "1L" Recurrent Neural Networks and CNN Network

This section examines light-weight architectures. In the initial experiment, convolutional neural networks (CNNs) were tested. Then, experiments were conducted with architectures based on single-layer recurrent neural networks (RNNs).

The recognition accuracy curves of each model was examined and compared in Figure 3. These steps were primarily designed to test the accuracy of the models that were developed using the provided dataset.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2024 doi:10.20944/preprints202410.2156.v1

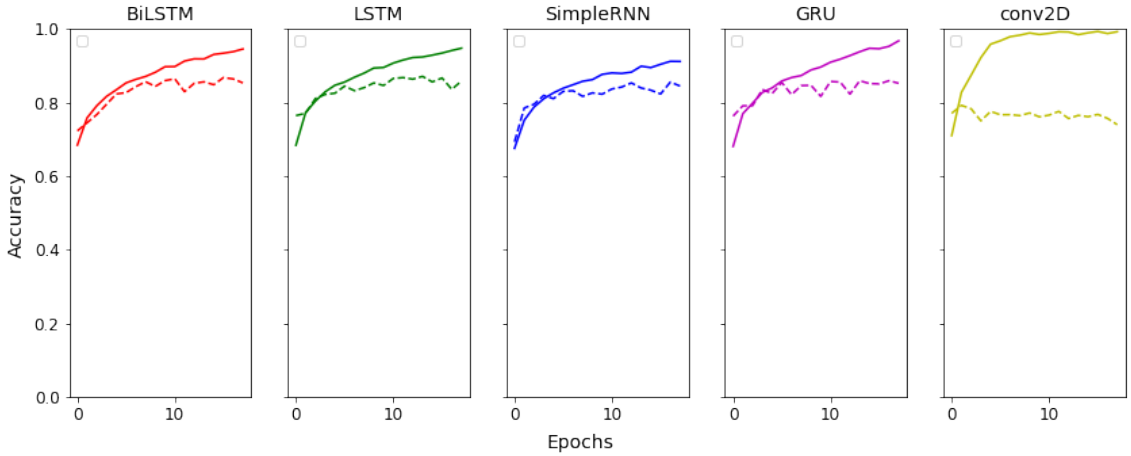


Figure 3. Confusion Matrix of deep learning model based on CNN network

The first experiment examined the performance of the CNN network, Figure 4, then the types of RNNs, Figure 5. For each model, the overall recognition accuracy and recognition curves were evaluated, and then the confusion matrices obtained from the validation data were examined. Although the confusion matrices are presented as the number of audio files, the overall classification results in percentage are presented in Table 11.

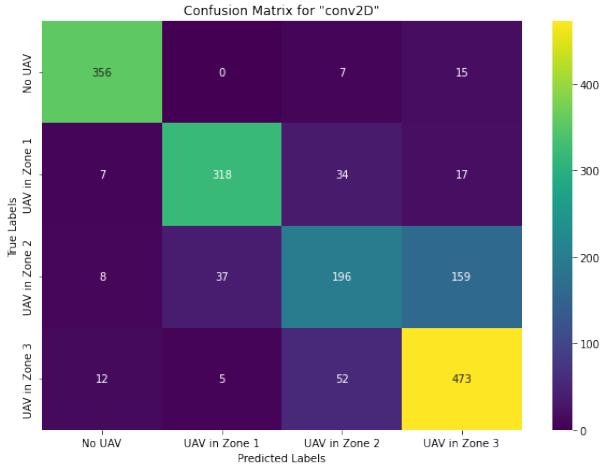


Figure 4. Accuracy curves of deep learning models based on "1L" recurrent neural networks and CNN network

 doi:10.20944/preprints202410.2156.v1

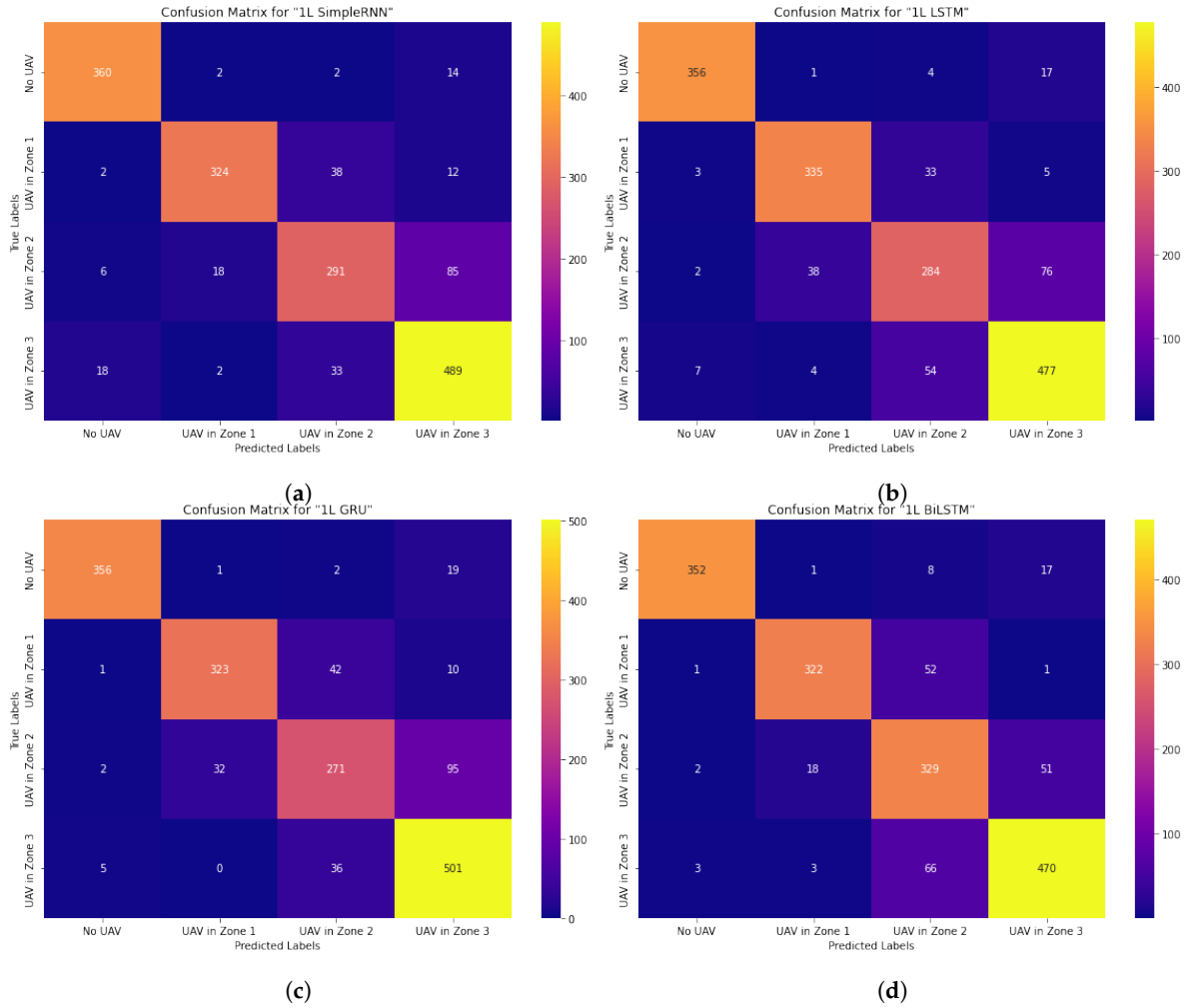


Figure 5. Classification Reports of Deep Learning models based on 1L recurrent neural networks

The CNN network showed less reliable results in estimating UAV distances during experimental tests at longer ranges. And the accuracy curve of the CNN network is also unrepresentative. In addition, the LSTM, BiLSTM, and GRU network models showed relatively good recognition capabilities among RNN-based architectures. In the following section, the voting approach was used to continue the tests, since none of the models could achieve a high level of reliable recognition alone for UAV range prediction tasks.

A total of five separate experimental studies were conducted. The results of each experiment were examined using the classification reports. In particular, the Precision, Recall, and F1-score metrics for each class were used to evaluate the results of each experiment. These metrics are used to objectively compare the quality of studies and determine their effectiveness. A thorough evaluation of the classification capabilities of each model can be demonstrated in the precision, recall, and F1-score statistics in Table 11.

Because, for all positively predicted objects, precision is the percentage of predicted positive results that are actually true-positive. Its formula is as follows:

$$Precision = \frac{T_p}{T_p + F_p} \tag{1}$$

 doi:10.20944/preprints202410.2156.v1

And Recall is the ratio of all true-positive objects to all positively predicted objects; it indicates the number of samples of all positive examples that were correctly classified.

$$Recall = \frac{T_p}{T_p + F_n} \tag{2}$$

The F1-score metric, which integrates precision and recall information, might be computed using these two metrics.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

Thus, Precision, Recall and F1-Score were calculated to comprehensively evaluate the classification ability of each model. The results of each test were compared using these measures to determine the best model or set of parameters. For example, high recall means that the model reduces false negatives, while high precision means that there are fewer false positives. The F1-Score takes into account the diversity of the data and evaluates how well these two criteria are balanced.

Table 11. The performance of the models and their prediction metrics.

Model	Classes	Precision, %	Recall, %	F1-Score, %
"2L" CNN				
	No UAV	93	94	94
	UAV in Zone 1	88	85	86
	UAV in Zone 2	68	49	57
	UAV in Zone 3	71	87	78
"1L" SimpleRNN				
	No UAV	93	95	94
	UAV in Zone 1	94	86	90
	UAV in Zone 2	80	73	76
	UAV in Zone 3	81	90	86
"1L" LSTM				
	No UAV	97	94	95
	UAV in Zone 1	89	89	89
	UAV in Zone 2	76	71	73
	UAV in Zone 3	83	88	85
"1L" BiLSTM				
	No UAV	98	93	96
	UAV in Zone 1	94	86	89
	UAV in Zone 2	72	82	77
	UAV in Zone 3	87	87	87
"1L" GRU				
	No UAV	98	94	96
	UAV in Zone 1	91	86	88
	UAV in Zone 2	77	68	72
	UAV in Zone 3	80	92	86

The results showed that the CNN showed low performance in detecting the presence of drones in high background noise conditions. The CNN network was unable to clearly distinguish between the audio distance data of the second and third zones. The main reason for this shortcoming is the significant presence of noise in the audio data recorded during the second zone. In particular, most of the data in this region was recorded with a high level of background noise. In addition, at long distances, the CNN showed relatively unreliable results in distance recognition, which may be due to the increasing influence of close-range noise from other objects as the distance increases.

Although SimpleRNN recurrent neural network (RNN) has the ability to distinguish the second zone relatively more accurately, the accuracy remains insufficient. In particular, although the SimpleRNN network provided satisfactory recognition results for the second zone, it was observed that in the validation data obtained during the real-time system study, it often made errors in the "no

 doi:10.20944/preprints202410.2156.v1

drone" class and the confusion matrix of the third zone data. Therefore, this model was not pursued for further study.

Moreover, the LSTM and GRU networks showed higher recognition accuracy for the second zone. In particular, the LSTM model was shown to be able to recognize objects better than the GRU network despite the influence of background noise. And the GRU network, on the contrary, had the ability to more reliably recognize objects at a long distance compared to LSTM. The BiLSTM network also demonstrated the ability to determine the drone's distance in a noisy background more accurately and reliably than the other two networks. However, the BiLSTM network showed weaker results than the GRU network in terms of recognizing long-distance objects. Therefore, the next section will provide a thorough analysis of the problem of combining the abilities of GRU, LSTM, and BiLSTM recurrent networks in the investigation using trained models based on the voting method.

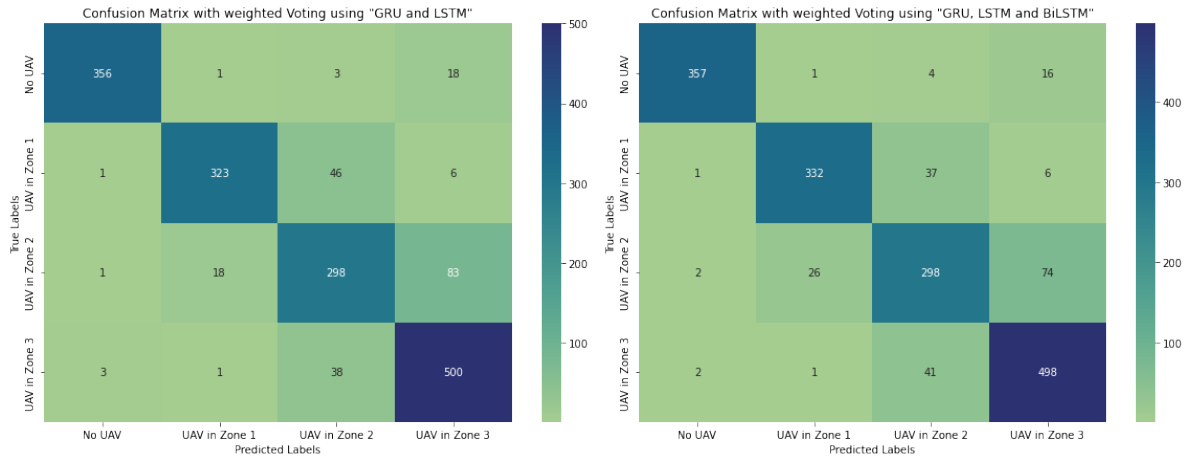
4.1.2. Prediction Based on Weighted Voting System Using Recurrent Neural Networks "1L"

In this section of the study, the capabilities of neural networks that were more robust in the previous section were combined using a voting strategy. One effective method for combining the predictions of multiple models into a single result in machine learning is a voting system based on trained models. By using this approach, it is possible to maintain the strengths of each model while minimizing their weaknesses. The overall model accuracy should be improved by comparing the results of different replicated networks.

The study used *Weighted Voting System* involving two different combinations:

1) Merging of GRU and LSTM networks, Figure 6a.

2) Merging of GRU, LSTM and BiLSTM networks, Figure 6b.



(**a**) (**b**)

Figure 6. Classification Reports of Voting System Using Recurrent Neural Networks "1L".

The reason for combining the trained models of GRU and LSTM networks is that the LSTM network is less confusing in the closest zone on Confusion Matrix than the BiLSTM network. The GRU network was relatively robust at long altitudes.

In our case, the priority weights for our models on the fisrt combination were assigned as follows:

M1 (weight for the model GRU) = 0.4, and M2 (weight for the model LSTM = 0.6).

The priority weights for the models on the second combination were assigned as follows:

M1) weight for the model GRU = 0.35;

M2) weight for the model LSTM = 0.45;

M3) weight for the model BiLSTM = 0.2;

Here, M1-M3 are the trained models. The model weights can be adjusted according to their individual recognition abilities. That is, the weighted voting system allows each model to be adjusted to participate in the decision-making process according to its reliability and performance.

 doi:10.20944/preprints202410.2156.v1

In our case, the combination of GRU, LSTM, BiLSTM gave a better recognition result from the analysis of two indicators, Figure 6 and Table 12.

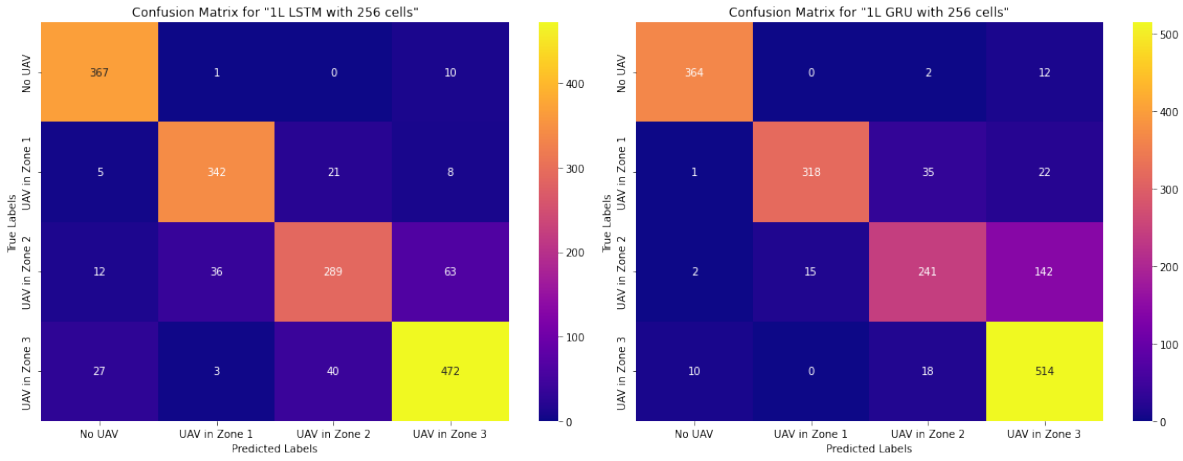
Table 12. The performance of the models and their prediction metrics.

Model	Classes	Precision, %	Recall, %	F1-Score, %
Voting System with "GRU-LSTM"				
	No UAV	99	94	96
	UAV in Zone 1	94	86	90
	UAV in Zone 2	77	74	76
	UAV in Zone 3	82	92	87
Voting System with "GRU-LSTM-BiLSTM"				
	No UAV	99	**94**	96
	UAV in Zone 1	92	**88**	90
	UAV in Zone 2	78	**74**	76
	UAV in Zone 3	84	**92**	88

4.1.3. Deep Learning Models Based on "1L" Recurrent Neural Networks "GRU" and "LSTM" with More Cells

The models' insufficient recognition accuracy motivated more studies with LSTM and GRU networks using a greater number of cells, Figure 7.

Here, although the classification responses show a reliable recognition rate, the confusion matrix increased the confusion with the non-drone class and the confusion between the zones. Therefore, the study of these two models was discontinued, Figures 7 and Table 13.



(**a**) "1L" LSTM with 256 cells (**b**) "1L" GRU with 256 cells

Figure 7. Classification Reports of "1L" recurrent neural networks "GRU" and "LSTM" with more cells.

Table 13. The performance of the models and their prediction metrics.

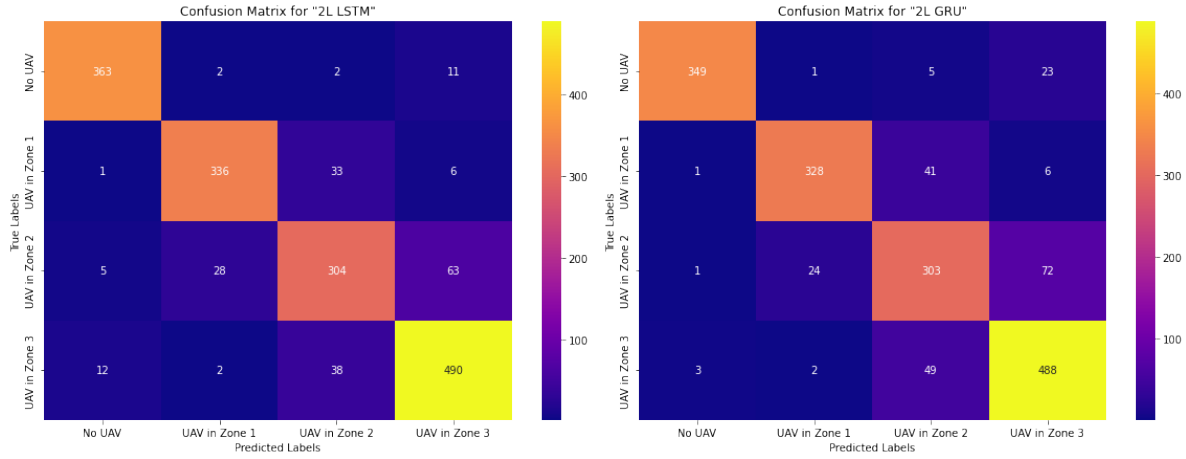
Model	Classes	Precision, %	Recall, %	F1-Score, %
"1L GRU" with 256 cells				
	No UAV	97	96	96
	UAV in Zone 1	95	85	90
	UAV in Zone 2	81	60	69
	UAV in Zone 3	74	95	83
"1L LSTM" with 256 cells				
	No UAV	89	**97**	93
	UAV in Zone 1	90	**91**	90
	UAV in Zone 2	83	**72**	77
	UAV in Zone 3	85	**87**	86

 doi:10.20944/preprints202410.2156.v1

4.1.4. Deep Learning Models Based on "2L" Stacked Recurrent Neural Networks "GRU" and "LSTM"

In this category of research, two layers of RNN were stacked with fewer cells at a time to carry out experimental work. With this approach, we can determine the impact of different architectures and parameters on the model and try to conduct a thorough evaluation of their effectiveness. To prevent overfitting problems and minimize computational costs, it is essential to reduce the number of cells in the layers.

These experiments showed that the stacked architecture is more efficient than increasing the number of cells. However, the LSTM network is still more confusing, Figure 8 and Table 14. Here, And LSTM and GRU networks gave similar recognition accuracies. Although it gave higher recognition accuracy than the results of previous research-oriented architectures, confusion was still encountered.



(**a**) "2L" LSTM (**b**) "2L" GRU

Figure 8. Classification Reports of "2L" recurrent neural networks "GRU" and "LSTM".

Table 14. The performance of the models and their prediction metrics.

Model	Classes	Precision, %	Recall, %	F1-Score, %
"2L GRU" with 64 cells				
	No UAV	99	92	95
	UAV in Zone 1	92	87	90
	UAV in Zone 2	76	76	76
	UAV in Zone 3	83	90	86
"2L LSTM" with 64 cells				
	No UAV	95	**96**	96
	UAV in Zone 1	91	**89**	90
	UAV in Zone 2	81	**76**	78
	UAV in Zone 3	86	**90**	88

4.1.5. Deep Learning Models Based on Hybrid Architectures

This section discusses the study of hybrid models. First, LSTM and GRU networks are hybridized. Then, they are hybridized by first providing GRU and then LSTM, Figure 9.

 doi:10.20944/preprints202410.2156.v1

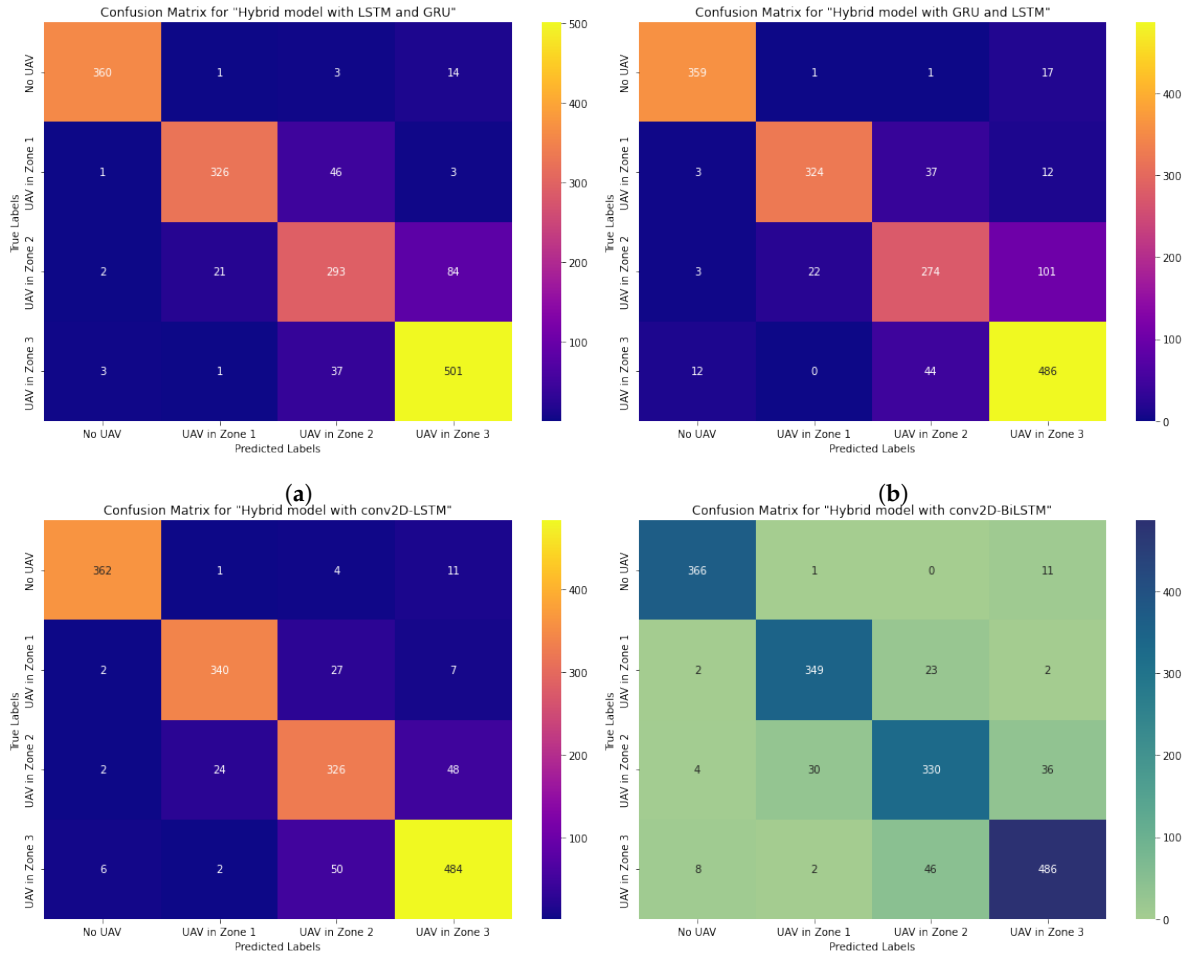


Figure 9. Classification Reports of Deep Learning models based on Hybrid Architectures.

Then BiLSTM and GRU networks were studied together. Then, using combinations with CNN networks, the LSTM and BiLSTM networks were evaluated individually. The combination of three networks, including CNN, BiLSTM, and GRU, was then taken into consideration.

The results of testing hybrid models show that the combination of CNN and BiLSTM networks is more effective. That is, 90% recognition accuracy for the farthest distance, 82% for the noisiest and most complex audio data in the middle zone, 93% for the closest zone, and 97% more accurate recognition of background noise.

The main goal of this category of experiments was to compare the pattern recognition skills of different architectures to assess how well they perform. The results of the experimental study aimed to compare model recognition abilities to determine which architecture perform best in specific applications.

 doi:10.20944/preprints202410.2156.v1

Table 15. The performance of the models and their prediction metrics.

Model	Classes	Precision, %	Recall, %	F1-Score, %
Hybrid model "LSTM-GRU"				
	No UAV	98	95	97
	UAV in Zone 1	93	87	90
	UAV in Zone 2	77	73	75
	UAV in Zone 3	83	92	88
Hybrid model "GRU-LSTM"				
	No UAV	95	95	95
	UAV in Zone 1	93	86	90
	UAV in Zone 2	77	69	72
	UAV in Zone 3	79	90	84
Hybrid model "BiLSTM-GRU"				
	No UAV	95	97	96
	UAV in Zone 1	92	87	90
	UAV in Zone 2	79	73	76
	UAV in Zone 3	84	91	87
Hybrid model "Conv2D-LSTM"				
	No UAV	97	96	97
	UAV in Zone 1	93	90	92
	UAV in Zone 2	80	81	81
	UAV in Zone 3	88	89	89
Hybrid model "Conv2D-BiLSTM-GRU"				
	No UAV	98	96	97
	UAV in Zone 1	92	91	91
	UAV in Zone 2	83	76	79
	UAV in Zone 3	85	92	88
Hybrid model "Conv2D-BiLSTM"				
	No UAV	96	**97**	97
	UAV in Zone 1	91	**93**	92
	UAV in Zone 2	83	**82**	83
	UAV in Zone 3	91	**90**	90

5. Discussions

In this study, the performance of different deep learning architectures for the audio identification system of unmanned aerial vehicles (UAVs) at different distances was investigated using empirical tests. Specifically, five directions were investigated: convolutional neural networks (CNNs); single-layer recurrent neural networks (RNNs); two-layer RNNs (stacking 2-layer RNNs); single-layer RNNs with boosted cells; a voting system using single-layer RNNs; and hybrid models. All these architectures were investigated to analyze how well they can predict the audio activity of UAVs at different flight distances. The study focused on an efficient architecture for processing audio data and predicting UAV movements when they are in different range zones, Figure 10. In our case, there were three zones where suspicious UAV movements were expected:

1) Zone 1 – closest zone (up to 15 meters);

2) Zone 2 – middle zone (from 15 to 30 meters);

3) Zone 3 – far zone (30-50 meters).

Figure 10 shows the validation of the real-time system, which was analyzed effectively, i.e., the real-time system with the CNN-BiLSTM model.

 doi:10.20944/preprints202410.2156.v1

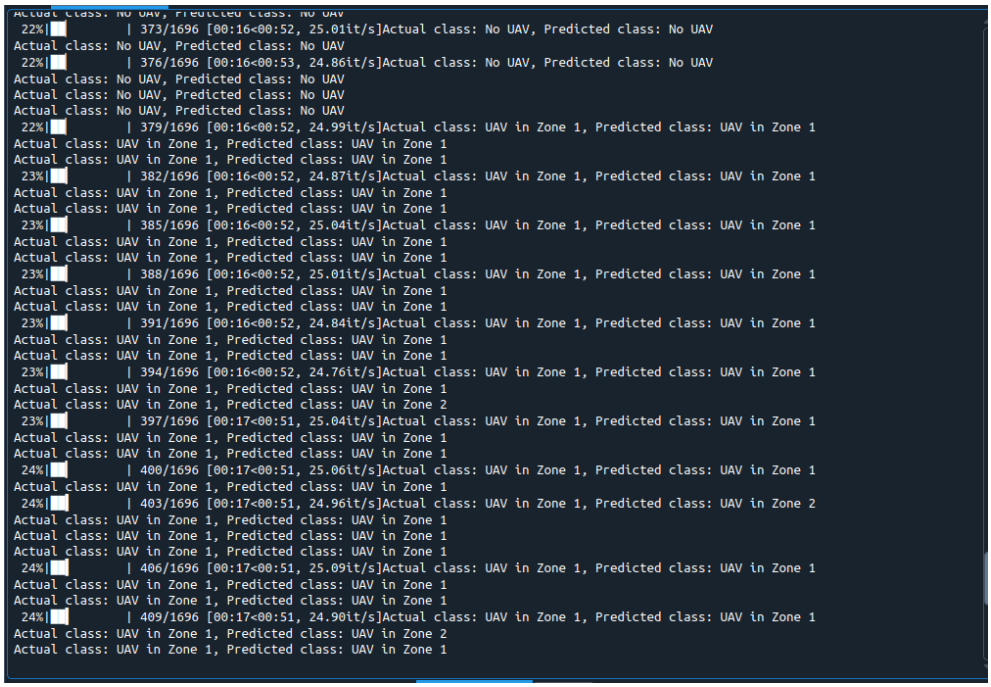


Figure 10. Prediction Results of the model CNN-BiLSTM.

And the UAV flight tests were conducted in various modes, including fully dynamic, partially static and semi-dynamic motion. It was found that the average speed of the UAV during the tests was between 3 and 15 meters per second. Based on this data, the maximum range of each zone was set. And the maximum range of the detection system was 50 meters. To verify the accuracy of the system in classifying UAV audio signals in real time, it was trained and tested in various flight modes. All the main measures for categorizing multi-class problems, including precision, recall and F1 score, were used in the validation phase.

As a result, as part of our research efforts, we have thoroughly analyzed several deep learning architectures for UAV audio data at different distances based on real-world tests. In summary, our work has contributed to the following solutions:

- The system's ability to determine and predict the distance to the UAV in real time has been thoroughly studied;

- Various deep learning architectures have been comprehensively studied.

So, it can be observed that the UAV distance prediction system based on deep learning is feasible and is found to work well with hybrid models, but more data is needed to apply this system in real-time conditions.

6. Conclusions

In conclusion, we tried to review a number of deep learning architectures for a real-time UAV audio detection system. The main objective was to determine which deep learning architectures perform best in predicting UAV audio data at different flight distances. The ability of CNN and RNN architectures to recognize audio inputs in the dynamic mobility of UAVs was analyzed as an effective solution.

The results showed that the CNN-BiLSTM architecture has relatively good prediction accuracy and performance. Similar to previous studies, the remaining RNN models showed robustness in UAV load estimation and binary classification tasks. However, the main drawback of our study was the lack of acoustic data, including many different new UAV models.

Despite this, the study showed that UAV ranges can be successfully predicted in real time using audio data when sufficient audio data is available and in moderately noisy regions. This confirms the potential of using the system as an additional acoustic sensor for protected area monitoring systems.

 doi:10.20944/preprints202410.2156.v1

In future, our research can be focused on solving two key problems: Increasing the amount of audio data for more accurate recognition and Developing bimodal systems that can combine audio with other sensors.

Thus, this study lays the foundation for further development and improvement of real-time acoustic monitoring systems for UAVs.

Author Contributions: Conceptualization, D.U., L.I., U.S., and E.M.; methodology, D.U.; software, D.U.; validation, D.U., L.I. and E.M.; formal analysis, D.U., and U.S.; investigation, D.U.; resources, D.U. and U.S.; data curation, D.U. and A. Y.; writing—original draft preparation, D.U.; writing—review and editing, D.U.; visualization, D.U., and S.U.; supervision, E.M. and L.I. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the Scientific Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant IRN AP14971907, "Development of a robust frequency-based detection system for suspicious UAVs using SDR and acoustic signatures"). The conclusions drawn here are the sole responsibility of the authors.

Data Availability Statement:

The materials and work will be available at this link. Alternately, a request can also be made by sending an email to dana.utebaieva@gmail.com.

Acknowledgments: We would like to sincerely thank Mr. Azamat Umbetaliyev for his essential help and support in collecting data from UAVs at specific distances. Additionally, we are grateful to Mr. Seth Adams for his deep learning tutorials on audio classification.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, S.; Kim, H.; Lee, S.; Gallagher, J.C.; Kim, D.; Park, S.; Matson, E.T. Convolutional Neural Networks for Analyzing Unmanned Aerial Vehicles Sound. 2018 18th International Conference on Control, Automation and Systems (ICCAS), 2018, pp. 862–866.
2. Bowman, E. Canada beats New Zealand in women's soccer as Olympic spy drone scandal grows. WVIA Radio, July 25, 2024.
3. Kesteloo, H. Drone Drama at Olympics: Canada Accused of Spying on New Zealand Soccer Team. DroneXL, July 24, 2024.
4. Man Arrested for Flying Drone over the SpaceX Facility. C-UAS Hub, 2024.
5. Seidaliyeva, U.; Akhmetov, D.; Ilipbayeva, L.; Matson, E.T. Real-Time and Accurate Drone Detection in a Video with a Static Background. *Sensors* **2020**, *20*. doi:10.3390/s20143856.
6. Seidaliyeva, U.; Ilipbayeva, L.; Taissariyeva, K.; Smailov, N.; Matson, E.T. Advances and Challenges in Drone Detection and Classification Techniques: A State-of-the-Art Review. *Sensors* **2024**, *24*. doi:10.3390/s24010125.
7. Utebayeva, D.; Ilipbayeva, L.; Matson, E.T. Practical Study of Recurrent Neural Networks for Efficient Real-Time Drone Sound Detection: A Review. *Drones* **2023**, *7*. doi:10.3390/drones7010026.
8. Utebayeva, D.; Yembergenova, A. Study a deep learning-based audio classification for detecting the distance of UAV. 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), 2024, pp. 1–7. doi:10.1109/EAIS58494.2024.10569107.
9. Lim, D.; Kim, H.; Hong, S.; Lee, S.; Kim, G.; Snail, A.; Gotwals, L.; Gallagher, J.C. Practically Classifying Unmanned Aerial Vehicles Sound Using Convolutional Neural Networks. 2018 Second IEEE International Conference on Robotic Computing (IRC), 2018, pp. 242–245. doi:10.1109/IRC.2018.00051.
10. Nijim, M.; Mantrawadi, N. Drone classification and identification system by phenome analysis using data mining techniques. 2016 IEEE Symposium on Technologies for Homeland Security (HST), 2016, pp. 1–5. doi:10.1109/THS.2016.7568949.
11. Sedunov, A.; Haddad, D.; Salloum, H.; Sutin, A.; Sedunov, N.; Yakubovskiy, A. Stevens Drone Detection Acoustic System and Experiments in Acoustics UAV Tracking. 2019 IEEE International Symposium on Technologies for Homeland Security (HST), 2019, pp. 1–7. doi:10.1109/HST47167.2019.9032916.
12. Cheranyov, A.; Dukhan, E. Methods of Detecting Small Unmanned Aerial Vehicles. 2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), 2021, pp. 0218–0221. doi:10.1109/USBEREIT51232.2021.9455043.

13. Papa, U.; Del Core, G.; Giordano, G.; Ponte, S. Obstacle detection and ranging sensor integration for a small unmanned aircraft system. 2017 IEEE International Workshop on Metrology for AeroSpace (MetroAeroSpace), 2017, pp. 571–577. doi:10.1109/MetroAeroSpace.2017.7999533.
14. Jang, B.; Seo, Y.; On, B.; Im, S. Euclidean distance based algorithm for UAV acoustic detection. 2018 International Conference on Electronics, Information, and Communication (ICEIC), 2018, pp. 1–2. doi:10.23919/ELINFOCOM.2018.8330557.
15. Kim, J.; Park, C.; Ahn, J.; Ko, Y.; Park, J.; Gallagher, J.C. Real-time UAV sound detection and analysis system. 2017 IEEE Sensors Applications Symposium (SAS), 2017, pp. 1–5. doi:10.1109/SAS.2017.7894058.
16. Jiqing, L.; Husheng, F.; Qin, Y.; Chunhua, Z. Quad-rotor UAV Audio Recognition Based on Mel Spectrum with Binaural Representation and CNN. 2021 International Conference on Computer Engineering and Application (ICCEA), 2021, pp. 285–290. doi:10.1109/ICCEA53728.2021.00063.
17. Yang, B.; Matson, E.T.; Smith, A.H.; Dietz, J.E.; Gallagher, J.C. UAV detection system with multiple acoustic nodes using machine learning models. 2019 Third IEEE international conference on robotic computing (IRC). IEEE, 2019, pp. 493–498.
18. Anwar, M.Z.; Kaleem, Z.; Jamalipour, A. Machine learning inspired sound-based amateur drone detection for public safety applications. *IEEE Transactions on Vehicular Technology* **2019**, *68*, 2526–2534.
19. Fagiani, F.R.E. Uav detection and localization system using an interconnected array of acoustic sensors and machine learning algorithms. Master's thesis, Purdue University, 2021.
20. Ahmed, C.A.; Batool, F.; Haider, W.; Asad, M.; Hamdani, S.H.R. Acoustic Based Drone Detection Via Machine Learning. 2022 International Conference on IT and Industrial Technologies (ICIT). IEEE, 2022, pp. 01–06.
21. Tejera-Berengue, D.; Zhu-Zhou, F.; Utrilla-Manso, M.; Gil-Pita, R.; Rosa-Zurera, M. Acoustic-Based Detection of UAVs Using Machine Learning: Analysis of Distance and Environmental Effects. 2023 IEEE Sensors Applications Symposium (SAS). IEEE, 2023, pp. 1–6.
22. Salman, S.; Mir, J.; Farooq, M.T.; Malik, A.N.; Haleemdeen, R. Machine learning inspired efficient audio drone detection using acoustic features. 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST). IEEE, 2021, pp. 335–339.
23. Ohlenbusch, M.; Ahrens, A.; Rollwage, C.; Bitzer, J. Robust drone detection for acoustic monitoring applications. 2020 28th European Signal Processing Conference (EUSIPCO). IEEE, 2021, pp. 6–10.
24. Solis, E.R.; Shashev, D.V.; Shidlovskiy, S.V. Implementation of audio recognition system for unmanned aerial vehicles. 2021 International Siberian Conference on Control and Communications (SIBCON). IEEE, 2021, pp. 1–8.
25. Lee, H.; Han, S.; Byeon, J.I.; Han, S.; Myung, R.; Joung, J.; Choi, J. CNN-Based UAV Detection and Classification Using Sensor Fusion. *IEEE Access* **2023**, *11*, 68791–68808. doi:10.1109/ACCESS.2023.3293124.
26. Ku, I.; Roh, S.; Kim, G.; Taylor, C.; Wang, Y.; Matson, E.T. UAV Payload Detection Using Deep Learning and Data Augmentation. 2022 Sixth IEEE International Conference on Robotic Computing (IRC), 2022, pp. 18–25. doi:10.1109/IRC55401.2022.00009.
27. Nakngoen, N.; Pongboriboon, P.; Inthanop, N.; Akharachaisirilap, J.; Woodward, T.; Teerasuttakorn, N. Drone Classification Using Gated Recurrent Unit. IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society, 2023, pp. 1–4. doi:10.1109/IECON51785.2023.10312193.
28. Solis, E.R.; Shashev, D.V.; Shidlovskiy, S.V. Implementation of Audio Recognition System for Unmanned Aerial Vehicles. 2021 International Siberian Conference on Control and Communications (SIBCON), 2021, pp. 1–8. doi:10.1109/SIBCON50419.2021.9438906.
29. Racinskis, P.; Arents, J.; Greitans, M. (POSTER) Drone Detection and Localization Using Low-Cost Microphone Arrays and Convolutional Neural Networks. 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), 2023, pp. 80–82. doi:10.1109/DCOSS-IoT58021.2023.00023.
30. Kim, B.; Jang, B.; Lee, D.; Im, S. CNN-based UAV Detection with Short Time Fourier Transformed Acoustic Features. 2020 International Conference on Electronics, Information, and Communication (ICEIC), 2020, pp. 1–3. doi:10.1109/ICEIC49074.2020.9051099.
31. Jeon, S.; Shin, J.W.; Lee, Y.J.; Kim, W.H.; Kwon, Y.; Yang, H.Y. Empirical study of drone sound detection in real-life environment with deep neural networks. 2017 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 1858–1862. doi:10.23919/EUSIPCO.2017.8081531.

32. Al-Emadi, S.; Al-Ali, A.; Al-Ali, A. Audio-Based Drone Detection and Identification Using Deep Learning Techniques with Dataset Enhancement through Generative Adversarial Networks. *Sensors* **2021**, *21*. doi:10.3390/s21154953.
33. Casabianca, P.; Zhang, Y. Acoustic-Based UAV Detection Using Late Fusion of Deep Neural Networks. *Drones* **2021**, *5*. doi:10.3390/drones5030054.
34. İlhan Aydın.; Kızılay, E. Development of a new Light-Weight Convolutional Neural Network for acoustic-based amateur drone detection. *Applied Acoustics* **2022**, *193*, 108773. https://doi.org/10.1016/j.apacoust.2022.108773.
35. Dumitrescu, C.; Minea, M.; Costea, I.M.; Cosmin Chiva, I.; Semenescu, A. Development of an acoustic system for UAV detection. *Sensors* **2020**, *20*, 4870.
36. Vemula, H.C. Multiple drone detection and acoustic scene classification with deep learning. Master's thesis, Wright State University, 2018.
37. Wang, Y.; Chu, Z.; Ku, I.; Smith, E.C.; Matson, E.T. A large-scale uav audio dataset and audio-based uav classification using cnn. 2022 Sixth IEEE International Conference on Robotic Computing (IRC). IEEE, 2022, pp. 186–189.
38. Katta, S.S.; Nandyala, S.; Viegas, E.K.; AlMahmoud, A. Benchmarking audio-based deep learning models for detection and identification of unmanned aerial vehicles. 2022 Workshop on Benchmarking Cyber-Physical Systems and Internet of Things (CPS-IoTBench). IEEE, 2022, pp. 7–11.
39. Taha, B.; Shoufan, A. Machine learning-based drone detection and classification: State-of-the-art in research. *IEEE access* **2019**, *7*, 138669–138682.
40. Mu, X.; Lu, J.; Watta, P.; Hassoun, M.H. Weighted voting-based ensemble classifiers with application to human face recognition and voice recognition. 2009 International Joint Conference on Neural Networks, 2009, pp. 2168–2171. doi:10.1109/IJCNN.2009.5178708.
41. Moattar, M.H.; Homayounpour, M.M. A Weighted Feature Voting Approach for Robust and Real-Time Voice Activity Detection. *ETRI Journal* **2011**, *33*, 99–109, https://onlinelibrary.wiley.com/doi/pdf/10.4218/etrij.11.1510.0158 https://doi.org/10.4218/etrij.11.1510.0158.
42. Zhang, H.; Zhou, Y. A Neural Network-Based Weighted Voting Algorithm for Multi-Target Classification in WSN. *Sensors* **2024**, *24*. doi:10.3390/s24010123.