Ministry of Science and Higher Education of the Republic of Kazakhstan

SDU University

Assel Yembergenova

Research on a UAV Distance Prediction System Based on Acoustic Data and Deep Learning

THESIS

Presented in Partial Fulfilment for the

Degree of Master of Technical Science in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **Dana Utebayeva**

Kaskelen, June 2025

SDU University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty of Engineering and Natural Sciences

Assistant Professor, PhD Akhmedov Ramis

«13» 06 2025

Topic of the thesis:
Research on a UAV Distance Prediction System Based on Acoustic Data and Deep Learning

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science", SDU University

Head of Department Mukash Zhanar

Academic Supervisor Utebayeva Dana

Master student Yembergenova Assel

Kaskelen, 2025

Declaration

I certify that this thesis entitled "Research on a UAV Distance Prediction System Based on Acoustic Data and Deep Learning" is my own work and has been done under the supervision of Dr. Utebayeva Dana, Associate Professor of SDU University.
It has not been submitted before, either fully or partially, for the award of a degree or diploma at this institution or any institution. The sources of information used in the thesis have been duly acknowledged through proper references and citation.
I understand the rules and regulations against plagiarism and academic honesty and certify that the research abides by them in its entirety.

Assel Yembergenova

June, 2025

Acknowledgements

I extend my sincerest appreciation to my supervisor, Dana Utebayeva, for her constant guidance, assistance, and motivation during the research process. Her valuable inputs and critiques have played an integral role in shaping this research and leading it to fruition.
I am also truly grateful to Azamat Umbetaliev for his kind help with the operation of drones. His assistance in performing safe flight testing and data collection was instrumental to this research being successful.
I could not have completed this dissertation without the assistance and guidance of all of these people, and I am sincerely obliged to all of them.

Dedication

I dedicate my work to my dear family, whose unrelenting support, encouragement, and conviction in me have been my greatest asset in undertaking this journey.
To my parents, I thank you for the boundless love, sacrifices, and the ever-inspiring urge to reach higher heights.
To my mentors and friends, your kind words, motivation, and wise advice have kept me energized and motivated.
To everyone who has ever believed in me and trusted in my ability–this victory is just as much yours.

Abstract

Unmanned aerial vehicles (UAVs), also referred to as drones, have become increasingly popular in recent years, posing serious security and privacy issues. Concerns have been raised by their growing presence in public areas and civilian life as a result of incidents involving disturbances, privacy invasion, and unauthorised surveillance. This study aims to address these issues by creating an intelligent, sound-based system that can identify drones and determine how close they are to people or sensitive areas.

The primary objective of this study was to assess the viability of classifying drone distance based on sound emissions using deep learning models and audio signals. Three zones—Zone 1, Zone 2, and Zone 3—each denoting varying degrees of proximity—were created from the drone sounds. Convolutional neural networks (CNNs), bidirectional long short-term memory networks (BiLSTMs), and a hybrid CNN-BiLSTM model were among the deep learning models examined in the study.

With an average classification accuracy of 90%, the hybrid CNN-BiLSTM model outperformed the others. This model is very accurate at predicting drone distance zones because it successfully captured both spatial and temporal features from the audio recordings.

These results imply that drone detection systems can be greatly improved by combining deep learning with audio-based classification. Such systems could significantly increase responsiveness and accuracy in detecting unauthorised UAV activity when paired with other sensory inputs in bimodal or multimodal frameworks. All things considered, this research advances acoustic sensing technologies to protect critical infrastructure and public safety from the increasing threat of rogue drone usage.

Keywords: Drone Detection, Deep Learning, Acoustic Sensing, UAV Proximity Estimation, CNN-BiLSTM, Audio Classification

Аңдатпа

Ұшқышсыз ұшу аппараттары (ҰҰА), яғни дрондар, соңғы жылдары кеңінен таралып, қауіпсіздік пен құпиялыққа байланысты маңызды мәселелер туындатуда. Қоғамдық орындар мен азаматтық өмірде дрондардың көбеюі түрлі тәртіп бұзушылықтар, жеке өмірге қол сұғу және рұқсатсыз бақылау сияқты жағдайларға себеп болды. Бұл зерттеу осы мәселелерді шешу мақсатында дрондарды анықтап, олардың адамдарға немесе маңызды аймақтарға қаншалықты жақын екенін анықтай алатын дыбысқа негізделген ақылды жүйені құруға бағытталған.
Зерттеудің басты мақсаты – дрондардың шығаратын дыбыстары арқылы олардың арақашықтығын терең оқыту модельдері арқылы классификациялау мүмкіндігін бағалау. Дрон дыбыстары негізінде үш аймақ – "1-аймақ", "2-аймақ" және "3-аймақ" – жасалды, әрқайсысы дронның жақындығының әртүрлі деңгейін білдіреді. Зерттеу барысында конволюциялық нейрондық желілер (CNN), екі бағытты ұзақ-қысқа мерзімді жады желілері (BiLSTM) және олардың гибридтік CNN-BiLSTM моделі қарастырылды.
90% орташа классификация дәлдігімен CNN-BiLSTM гибридтік моделі басқа модельдерге қарағанда үздік нәтиже көрсетті. Бұл модель дыбыстық жазбалардан кеңістіктік және уақыттық ерекшеліктерді тиімді анықтау арқылы дронның арақашықтық аймағын дәл болжай алды.
Зерттеу нәтижелері дыбыстық классификация мен терең оқытуды біріктіру арқылы дрондарды анықтау жүйелерін едәуір жетілдіруге болатынын көрсетеді. Мұндай жүйелер екі немесе көпмодальды құрылымдарда басқа сенсорлық деректермен біріктірілгенде рұқсатсыз ҰҰА әрекеттерін жылдам және дәл анықтауға мүмкіндік береді. Жалпы алғанда, бұл зерттеу акустикалық сенсорлық технологияларды жетілдіре отырып, маңызды инфрақұрылым мен қоғамдық қауіпсіздікті рұқсатсыз дрондардан қорғауға үлес қосады.
Кілт сөздер: Дронды анықтау, Терең оқыту, Акустикалық сезу, ҰҰА қашықтығын бағалау, CNN-BiLSTM, Дыбысты жіктеу

Аннотация

Беспилотные летательные аппараты (БПЛА), также известные как дроны, в последние годы стали чрезвычайно популярными, что вызывает серьёзные проблемы безопасности и конфиденциальности. Их всё более широкое присутствие в общественных местах и в гражданской жизни вызывает обеспокоенность из-за инцидентов, связанных с нарушением порядка, вторжением в личную жизнь и несанкционированным наблюдением. Настоящее исследование направлено на решение этих проблем путём создания интеллектуальной системы, основанной на звуке, которая может выявлять дроны и определять их близость к людям или чувствительным зонам.

Основная цель исследования — оценить возможность классификации расстояния до дрона на основе звуковых сигналов с использованием моделей глубинного обучения. Звуки дронов были разделены на три зоны — «Зона 1», «Зона 2» и «Зона 3», каждая из которых обозначает различную степень близости. В исследовании были изучены различные модели глубинного обучения, включая сверточные нейронные сети (CNN), двунаправленные сети долгой краткосрочной памяти (BiLSTM) и гибридную модель CNN-BiLSTM.

Среди всех моделей гибридная CNN-BiLSTM показала наилучшие результаты, достигнув средней точности классификации 90%. Эта модель эффективно извлекала как пространственные, так и временные характеристики из аудиозаписей, что обеспечивало высокую точность в определении зоны расстояния до дрона.

Полученные результаты показывают, что системы обнаружения дронов можно значительно улучшить за счёт объединения глубинного обучения и звуковой классификации. При использовании в составе бимодальных или мультимодальных систем, такие технологии могут существенно повысить точность и оперативность в обнаружении несанкционированной активности БПЛА. В целом, данное исследование способствует развитию акустических сенсорных технологий для защиты критически важной инфраструктуры и общественной безопасности от растущей угрозы несанкционированного использования дронов.

Ключевые слова: Обнаружение дронов, Глубокое обучение, Акустический сенсор, Оценка расстояния БПЛА, CNN-BiLSTM, Аудио классификация

Abbreviations

UAV – Unmanned Aerial Vehicle

CNN – Convolutional Neural Network

RNN – Recurrent Neural Network

LSTM – Long Short-Term Memory

GRU – Gated Recurrent Unit

BiLSTM – Bidirectional Long Short-Term Memory

Hz – Hertz

m/s – Meters per Second

SNR – Signal-to-Noise Ratio

AI – Artificial Intelligence

DL – Deep Learning

ML – Machine Learning

STFT – Short-Time Fourier Transform

MFCC – Mel-Frequency Cepstral Coefficients

FPS – Frames Per Second

CPU – Central Processing Unit

GPU – Graphics Processing Unit

FFT – Fast Fourier Transform

RADAR – Radio Detection and Ranging

GPS – Global Positioning System

Table of Contents

Chapter 1

Introduction

The application of drones has grown a great deal in a variety of fields due to technological improvements, but with that, they have also raised the risks and incidents of their usage. Their expanded presence in military, recreational, commercial, and industrial fields and applications has generated innovation as well as concern. Now drones are widely utilized in fields like infrastructure inspection, logistics, agriculture, cinematography, and rescue operations. But with the growing accessibility and presence of drones, so does the threat of their abuse. Consequently, there has also been a rise in incidents with drone-induced crashes, unauthorized flying, and malicious attacks. To manage these challenges, researchers across the globe are exerting their best efforts in developing measures and solutions that will enhance the security and regulation of their application. The necessity of developing effective and smart drone detecting and monitoring mechanisms has thus developed significantly in the international research community [1].

Drones have become successively used as both legitimate and illegal instruments, frequently engaged in illegal missions posing serious threats to public and national security [2–4]. They appeal due to their affordability, ease of availability, and autonomous flight capabilities over long distances with limited supervision by the operator, making them popular not just with hobbyists and business organizations but also with malicious actors. Drones' versatility in operation has seen various forms, including illegal surveillance, data stealing, smuggling, and even drug delivery across borders. Of greater concern, in this case, is their increasing use in reconnaissance and intel-collection missions, where they collect confidential information by flying over forbidden or strategically sensitive areas. As observed in a variety of real-life incidents [2–7], drones have penetrated civilian airspaces above critical infrastructure, state buildings, military bases, and public events, exposing serious weaknesses in current aerial detection systems for threats. These incidents highlight the pressing need for developing counter-drone systems that not only detect but also classify and act against illegal UAV activity in real time. The possible outcomes in case drone incursions go unnoticed in situations can be far reaching, ranging from incursion into the nation's airspace, obstruction to public services and masses, to catastrophic loss in public events. It has thus become necessary for developing strong and scalable drone detection frameworks for protecting civilian and state interests as well as enhancing overall security. These incidents proved that if not deterred in time, drones can serve as catalysts for high-hazard incidents, and hence, advocating for the introduction of efficient monitoring and mitigation systems is a critical aspect for modern security framework [8–10].

Recent incidents [2–4, 7] evidently demonstrate that drones have the potential to be extremely dangerous for critical sites and infrastructure. Aside from mere visual monitoring, drones have the capability for malicious actions, like stealing secret files or inflicting physical injury. This necessitates effective countermeasures for drone threats. For these, real-time intelligent systems play an essential part. These intelligent systems have the ability to predict the approach of drones from protected regions and promptly identify unauthorized flights. By doing so, they can serve as a security warning plan. For the implementation of effective protection, specialized bimodal or multimodal drone control mechanisms have to be utilized. It should be possible for them to predict a drone approaching the guarded regions and thus avoid impending security issues. Specifically, they should have the ability to predict the proximity of a drone from guarded regions or recognize suspicious flight patterns immediately. Such systems, through the implementation of more than one sense modality such as visual, sound, thermal, and radar, have the ability to provide more effective and timely detection services, particularly in environments where one modality may not be effective due to noise, obstruction, and range restrictions [11].

Sound localization has been largely utilized in sound-based unmanned aerial vehicle (UAV) surveillance systems as a method for approximating the position and trajectory of airborne targets using passive acoustic sensing methods [12–15]. These systems take advantage of the intrinsic acoustic signatures emitted by drones in flight, which are generated largely by the fast spinning of propellers and the motor operation. The central concept in acoustic localization is based upon detecting variations in the time of arrival (ToA), phase, and/or amplitude of sound waves at a number of space-separated microphones in a microphone array. By processing these inter-microphone variations, sophisticated signal processing schemes can ascertain the direction of arrival (DoA) of sound, and in some configurations, also make estimates of the range or proximity to the UAV. This spatial data, once determined, contributes actionable information that facilitates the tracking and possible intercepting of unwanted drones, even in cases where visual or radar-based methods might be hampered by line-of-sight restrictions, adverse weather, or dirty backgrounds. Furthermore, the passive operation of sound localization systems gives them the added advantage of being stealthy and power-efficient, as they neither transmit signals nor draw power from the environment, instead collecting ambient sound energy. With drone activity expected to grow, particularly in sensitive or off-limits airspaces, incorporating sound localization units in UAV detection systems offers a viable area for increased situational awareness and enabling timely reaction capabilities for both civilian and military-focused surveillance systems [16–19].

There are numerous possibilities with recent advances in deep learning-assisted audio classification. Sound analysis-based acoustic intelligence systems are also under investigation for predicting distance from drones based on their sounds. This work tries to investigate the intelligent sound-based UAV prediction system for integrating bimodal or multimodal UAV detection systems. Deep learning algorithms, especially convolutional and recurrent neural networks, have demonstrated a great deal of promise for learning sophisticated patterns in sound from raw audio signals. It is possible for these models to be educated for classifying audio signals not just by model or by drone type but also by proximity, which is crucial for early detection and threat assessment. The proposed system will utilize these functionalities for supporting more extensive sensory systems, which must function reliably under real-time and under different environmental conditions [20].

Hence, the research sought to investigate deep learning models for predicting drone distance through classification. That is, the feasibility of using drone audio signals for predicting distance will be evaluated by comparatively examining different frameworks through experiments. For this, the following goals were devised:
1) An in-depth examination of deep learning model architectures for drone distance estimation from audio signals. This comprises examining how each model learns and generalizes useful acoustic features from audio created by drones.
2) To compare the performance of architectures like recurrent neural networks, convolutional neural networks, hybrid models and weighted vote systems for UAV distance prediction from sound. The experiments are meant to reveal the pros and cons of each method based on classification accuracy, generalization and computational speed. Therefore, this research sought to explore a multimodal or bimodal sensory structure based on deep learning models for predicting UAV distances from audio signals. By addressing the acoustics component in the context of a multi-level detection platform, the research provides the foundation for an intelligent and robust approach for tackling increasing threats from unauthorized and malicious UAV incursions into critical regions.

Chapter 2

Related works

Deep learning techniques have persistently proved their robustness and effectiveness in the field of drone sound detection and classification as indicated by a wide range of research evidence. Some of the most widely used and most notable of these neural network architectures include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and blended models that combine the powers of RNNs and CNNs. Deep learning architectures have proved effective in a wide range of activities such as, but not limited to the classification of drone types, flight state recognition, and binary and multiclass classification of drone activities [1, 5, 7, 20, 21].
In recent years, one of the arising and continually rising research thrusts in the area of drone detection and classification has been the use of drone-produced acoustic sounds for a variety of applications, including drone distance estimation from the point of detection and improving the accuracy and resilience against classification in real-world noise levels and interfering environmental conditions. These sound waves, generated by the motor and propellers of the drone in flight, hold important information that can be tapped into for building passive and non-intrusive surveillance systems. Using this form of audio data, however, is challenging in that it involves not a mere basic signal processing and/or some simple form of machine learning, but a number of different challenges that go beyond such tasks. Solving those requires not just tremendous progress in the design and development of proper model architectures, but also in those attendant fields like preprocessing strategies, advanced feature engineering methods, and practical requirements for real-time deployment. Specifically, data preprocessing plays a central role in ensuring that the input to models preserves the important acoustic features without excessive noise and environmental distortion. The feature engineering should be well-designed so as to extract those pertinent time-frequency patterns that differentiate drone noises from interfering sound waves or background noises. Central to this line of research is the development of deep network architectures specifically optimized for sequential processing of sound waves. Recurrent Neural Networks (RNNs), and a next-level version like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), as well as those based on attention, have become priority targets for researchers due to their far greater capacity for learning and capture of temporal dependencies within continuous sound sequences. These architectures lend themselves to building systems capable of detecting weak temporal cues and contextual patterns in drone sound, which is necessary for both this classification and distance estimation task. Consequently, they hold immense potential for making future breakthroughs in sound-based UAV detection systems. [22, 23].

One of the commonly applied and well-developed methods in Unmanned Aerial Vehicle (UAV) acoustical detection is sound source localization with Acoustic Sensor (AS) systems. This method entails approximating the space coordinates of a UAV by detecting and processing the time difference of arrival (TDOA) of sound coming from the UAV's motor and rotor through multiple strategically placed microphones in a geometric-array setup. The microphones can be configured in static geometric patterns for maximizing spatial resolution and accuracy in calculating the sound source's direction and distance. By computing the differences in time arrival for the same sound signal at every microphone, it would be possible to localize UAV's position with the help of signal processing algorithms, finally transforming time differences into space coordinates. This approach supports effective tracking and localization of UAVs in different environments with the specific advantage of being independent from visual sensors like cameras or active RF-based methods like radar. Since it is a passive detection system, it is especially beneficial in situations where visual impairment, RF noise, or stealth requirements exclude the application of conventional sensing mechanisms. Furthermore, the non-intrusive and low-cost aspect of sound sensing systems also increases the applicability for real-time UAV monitoring, particularly in the area of security, surveillance, and wildlife preservation. [12–19].

For instance, Nijim et al. were mainly interested in the identification and classification of drones from their sound patterns. Their approach utilized audio recordings of drones like DJI P3 and Quadcopter FPV 250 observed under different flight modes such as hover and active flight. Audio recordings were analyzed through the application of Hidden Markov Models (HMMs), clustering techniques, and data mining strategies for creating identifiable patterns and providing credible identification [24].

Another advanced system, the Drone Acoustic Detection System (DADS) proposed by researchers from Stevens Institute of Technology, made use of the acoustic signature of propellers for detecting, tracking, and classifying UAVs. Using an array of microphones, the system calculated the direction of arrival (DOA) and utilized triangulation algorithms for localizing the UAVs. Although the system successfully classified drones such as the DJI Phantom 4, Intel Falcon 8+, and DJI S1000 [25–27] under quiet conditions (up to 300 meters), its performance was significantly affected in noisy conditions, reflecting the limitations of real-world deployment [28].

Cheranyov et al. [29] looked into a multi-modal detection approach blending acoustics, radar, and optical detection. Their acoustics relied on microphone arrays for sound direction and angle estimation, hence UAV trajectory monitoring. While the range was constrained to around 15 meters, the approach was especially effective in detecting low-flying objects and coping with plastic material reflections that are prevalent on drones. Other advances have resulted from the application of low-cost sensors in UAV control. The work in tested Sonic Ranging Sensors (SRS) and InfraRed Sensors (IRS) under control of an Arduino Mega 2560 for calculation of distances and obstacle avoidance. Sensor readings were sampled 0.5 seconds apart and fused by means of algorithms for integrated data in order to minimize detection uncertainty. Although the system was effective for close-range detection, the application scope was limited to small quadcopters [30].

Researchers in one study suggested incorporating Short-Time Fourier Transform (STFT) with Euclidean Distance (ED) for detecting acoustics of drones. This method effectively differentiated between drone sounds and other ambient sounds based on the temporal frequency of drone audio. This comparatively simple yet effective method has promise for real-time monitoring for security-critical applications [31].

A number of recent works have presented new applications of deep learning and machine learning for detecting drones acoustically. Past works [5,20] have made their contributions by presenting lightweight models for UAV identification from Mel-spectrogram features and Recurrent Neural Networks. The work presented a method of UAV distance estimation by training a neural network based on a Gated Recurrent Unit on audio recordings captured at different altitudes (5–50 m). This method showed impressive classification success—94% for 10-meter ranges and 98% for 15-meter ranges. Even though the work has not yet been tested extensively over broader sets of data and under full-dynamic conditions, it sets a promising base for unifying the acoustic signals with the other senses in real-time UAV identification applications [20].

A comparison of several machine learning and deep learning models for drone sound classification was presented in [32]. The authors developed a Mel spectrogram-based CNN classifier and performed comparative assessments with other audio features like Mel Frequency Cepstral Coefficients (MFCCs) and classifiers like Support Vector Machines (SVMs). The experiments proved the excellence of Mel-spectrogram inputs based on CNN models regarding accuracy and noise robustness.

A new approach that integrated FFT with Plotted Image-Based Machine Learning (PIL) and k-Nearest Neighbors (KNN) was outlined in [33]. The system, which utilized real-time detection ability and was cost-effective, had a detection rate of 61%. Signal preprocessing and noise removal were indicated by the authors for further increasing the performance of the system.

A diverse array of other studies [34–41] has applied machine learning models for UAV sound recognition, highlighting the utility of features like MFCCs and spectral roll-off. Furthermore, studies such as [1, 21, 42–55] explored CNNs, RNNs, and their hybrid architectures for drone sound classification, experimenting with different hyperparameter tuning techniques and input features. Comprehensive literature reviews, like those found and have provided valuable insights into the capabilities and limitations of these approaches [7, 56].

In efforts to maximize classification robustness and accuracy, ensemble techniques like the method of weighted voting have come into focus for acoustic signal recognition. [57] studied ensemble classifiers in combination with a weighted voting approach for facial and vocal recognition. The approach proved effective in weighting the classifiers' results and greatly improved global recognition with higher weights allocated to more confident or correct models.

In the same manner, in [58], the researchers developed a low-complexity and robust voice activity detection scheme with a weighted voting approach. The scheme performed better than baseline schemes under conditions of channel distortion and poor threshold values, illustrating the worth of such a tactic for real-world signal classification.

A key contribution in this domain was from the research [59] that proposed a neural network-based weighted voting classifier. Experimentation proved that the ensemble method improved classification by around 5% in comparison with isolated neural networks. Yet, it also raised computational overhead and memory needs, and thus there was a trade-off between accuracy and efficiency. Besides that, several other studies have also confirmed the efficacy of using weighted voting schemes in other contexts [59–62]. All these studies have established that ensemble learning approaches, when backed by proper weighting schemes, are able to perform much better than individual model designs with regard to classification accuracy, particularly in noisy or complicated environments. Cumulatively, the literature reviewed herein explicitly suggests that deep learning meth-

ods coupled with weighted voting mechanisms—more specifically, using CNNs, RNNs, or their hybrids—are necessary for the design of efficient, trustworthy, and scalable drone detection tools. As such, the objective of this research is further to explore and analyze the combination of deep learning architectures with weighted voting techniques in the context of UAV audio signal identification [63].

Chapter 3

Methodology

3.1 Data collection

Deep convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have emerged as powerful, versatile instruments in the field of audio signal processing, especially for tasks that entail the identification of intricate patterns, for example, speech, environmental sound classification, and the identification of drones (unmanned aerial vehicles, or UAVs) [64,65]. These deep architectures found in both models are highly appreciated for their respective capabilities to extract various types of characteristics from raw or lightly processed sound waves—hierarchical spatial features in the case of CNNs, and sequential temporal patterns in the case of RNNs. CNNs perform well in detecting local spatial dependencies in time-frequency representations (e.g., spectrograms), recognizing patterns that enable discriminating between sound classes. In contrast, RNNs, especially those using higher-level variants like Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU), can represent temporal dynamics through retaining contextual cues in successive time steps, which is important for understanding sound evolution over time. Existing works proved that merging the two architectures—in joint or hybrid model configurations—tends to produce greater robustness and classification accuracy. This is primarily attributed to the complementary capabilities between the extraction of spatial vs. temporal features: as CNNs attend to the fine structure of sound frames, RNNs add value by representing the continuity and evolution in sound patterns. Consequently, hybrid application of CNN and RNN-based models has become a promising avenue in developing more mature, accurate, and generalizable sound recognition systems [66,67]. However, before such models of deep learning can be successful at such tasks of recognition, there is a necessary condition that first needs to be met—diligent and careful data preparation. This is particularly vital in the area of application for deep learning systems in audio-based recognitions, where quality, uniformity, and preprocessed input data control every aspect of model building. These conditions directly impact the training, the behavior of learning algorithms in converging, and, above all, the performance of the model upon completion and the capacity for generalization against unseen or real-world instances [68]. In contrast to structured tabular data, raw sound signals are inherently unstructured, noisy, and susceptible to variability caused by environmental factors, background noise, and variations in recording hardware. Therefore, sound signals must be preprocessed extensively before they can be ready for input into a neural

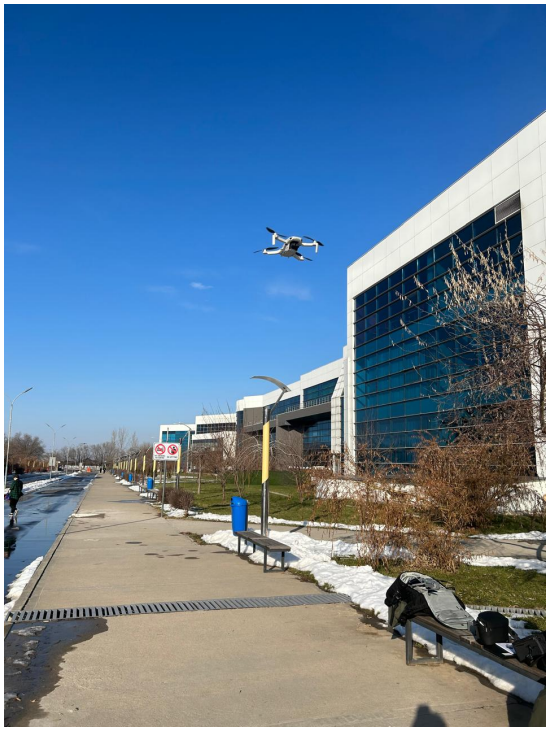
network. This preprocessed pipeline generally includes normalizing to provide constant amplitude scaling, noise removal to eliminate inaudible or confusing background noises, and segmentation to extract meaningful sections of the sound. In addition, sound needs to be translated into even more structured and interpretable representation forms, such as time-frequency representations like spectrogram, Mel-spectogram, and MFCCs, that can capture more effectively the necessary acoustic features to classify with accuracy. Moreover, data augmentation methods like time-stretching, time-stretching with pitch shifting, and SpecAugment also come in to augment robustness by exposing the model to greater variations in signals. Most of these preprocessed methods form the building block in ensuring that models based on learning in depth on sound data prove both functional and robust upon real-world application [69, 70].

Working with acoustic data generated with drones requires a collection of specific difficulties and complexities. Drone sounds, produced mainly by rotating propellers and motors, have acoustic characteristics that not only differ between different types of drones, but also between different regimes of flight, for example, during take-off, hovering, cruising, and landing. In addition, acoustic signals are typically recorded outdoors under uncontrolled conditions, where there can be many sources of background noise, including wind, traffic, human speech, animals, or other machinery. Environmental factors like temperature, humidity, ground surface, and obstacles (roof tops, trees, etc.) influence how sound propagates through the atmosphere and changes its timbre and intensity at different distances to the source. All this renders preprocessing and normalizing of drones' audio data all the more critical, as the neural network is to be trained on data that reflects well the variety of conditions it will see at deployment [71].

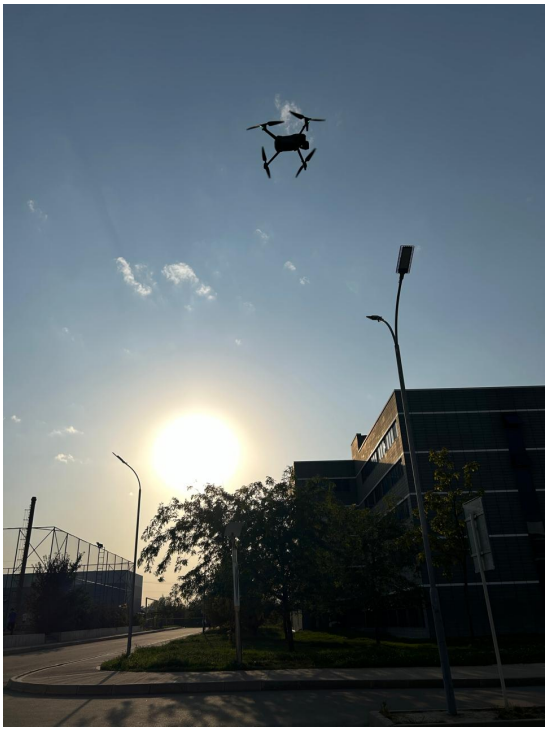
This section is devoted to an explanation of the data collection and data preprocessing process employed in this research, which involved training a neural network to classify or predict the distance of UAVs based on their sound profiles at different distances. In particular, the audio inputs to the training and test steps of deep learning models were recorded employing a ground-mounted Acoustic Sensor Point (ASP), as depicted in Figure 3.1. The ASP arrangement was critical to record sound data that realistically mimics what will be heard under realistic situations of surveillance and detection under practical scenarios. The drones were flown at different distances and heights—up to 50 meters—away from the ASP to mimic realistic conditions of surveillance and around critical infrastructure or safeguarded regions. For purposes of incorporating diversity and achieving generalizability to a large variety of UAVs, three different models of drones, i.e., the DJI Mini 2, Qazdrone, and DJI Air 3, as detailed in Table 3.1, were used for the experiment. These different models of UAV differ based on size, motor specification, propeller design, weight, and maximum range, all of which affect their inherent acoustic characteristics. Introducing this heterogeneity into the drone dataset was done with the intention of training the model to learn to be resilient to inter-model differences—a capability required for deployment to real environments where unknown or unauthorized UAV detection is usually a mandate.

The process of recording was well-planned and controlled. In every session, the microphone was kept at ground level, mimicking a fixed surveillance point or ground-level detection site. With this arrangement, an uninterrupted and clear line-of-sight of the drones was obtained as recordings were made at determined distances and elevations. With systematically varied positions of the UAV with respect to the ASP, a dataset containing samples of sounds at several segments, approaches, and orientations of drones was prepared. Each session was also recorded with accurate metadata of UAV model,

flight route, speed, atmospheric conditions, and time-stamp.
In addition, to validate that data obtained was of good quality and fit for model training, recordings went through a multi-phase preprocess pipeline. Audio signals were first sampled at a constant rate, and then were partitioned into equal-duration segments for simplified processing and batch training. Background noise removal was achieved through filtering methods like spectral subtraction and adaptive noise gating. The clean segments were subsequently converted into Mel-spectrogram representations—these being preferred as they have mimicking properties to simulate human auditory behavior and have been found to be effective for sound classification under deep neural networks. The last preprocessed dataset was subsequently partitioned into training, validation, and test sets, with all three types of UAVs being represented within these partitions. Particular care was needed to prevent data leakage through ensuring segments of the same flight as not being partitioned between different sets. Such methodological stringency during data preparation, as highlighted in this research, is crucial for developing deep learning models, which, although accurate in controlled environments, must be robust and trustworthy in dynamic and potentially noisy real-world environments.
Here, it was decided not to use publicly available datasets of UAV sounds. While there are publicly available datasets, these tend to be small and lack critical contextual details, especially accurate descriptions of the distance between the source and receiver. Because this research's main goal is to study how the acoustic signature of a UAV varies with distance and environmental characteristics, it was critical to capture custom audio data that are dense with distance-dependent features and contextual variety. The publicly available datasets, useful as these may be for generic classification of sounds or detection of UAVs, are not well suited for this study's fine-grained needs, which concern actual deployment of UAV detectors to field environments.
In order to fill this gap, a new data set was created through a sequence of well-designed and conducted field recordings. The collection process was carried out near a sports ground located beside a university campus, inside the university campus. It was selected for several reasons. First, it lies within a suburban region that presents a common realistic environment where UAVs may be operating or being detected — not too rural and not too urban. Second, the region hosts an average amount of background noise, providing an authentic environment with several non-UAV acoustic events, which are a challenge for proper recognition of UAVs.
Throughout the data collection campaign, the sessions of recordings were spread out several days apart to introduce variability to the background noise. The sessions were performed at various times of day to capture changes to human activity, car traffic, and environmental conditions including wind or bird sounds. During launch of the UAVs, during which time there was only sporadic acoustic data, the acoustic sensor—a sensitive microphone—was fixed to the ground at a point, known as the Acoustic Sensor Point (ASP). Fixed positioning facilitated consistent measurement of the acoustic footprint of the UAV at different distances. The UAVs were flown at incremental distances away from the ASP, to a maximum vertical distance of 50 meters, and at horizontal distances which provided for coverage to gather a variety of sound intensities and propagation characteristics. The chosen site had a dynamic soundscape. For instance, during certain sessions of recordings, there was evident background activity including, for instance, students playing football on nearby fields, campus maintenance activities, human communication, and vehicle movement within car parks of the university. Additionally, during certain days, there were increased presence of outside urban sounds including, for instance, sounds



(a)



(b)

(c)

Figure 3.1 – Data collection process

from mini-truck and delivery vehicles plying nearby roads, occasional horn blowing, or revving of engines. Such concurrent sound events made the dataset richer and therefore more applicable to practical purposes. More importantly, there were recordings where there existed concurrent background sounds with concurrent UAV flight, providing complicating acoustic interference that the neural network models must untangle.
A fifth class of ambient environmental noise was, moreover, recorded under the same conditions but without the presence of a UAV. It contained sounds of rustling vegetation, wind gusts of differing strengths, and chirping birds, and was an important inclusion for learning to discriminate between sounds produced by a UAV and those naturally found within the environment. Wind provided low frequencies, while swaying branches of trees contributed non-rhythmic patterns to the signal—a problem an efficaciously trained deep model must be able to cope with.
Additionally, since there is a variety of types of UAVs and acoustic profiles, three different models of UAVs were utilized during data collection: DJI Mini 2, Qazdrone, and DJI Air 3. All these have varying motor and propeller configurations, flight characteristics, and acoustic patterns, influenced differently by range and environmental interferences. The capture of these recordings was systematic, with each model being flown at specific distances and altitudes, permitting equivalence and consistent labeling and feature extraction.
In brief, building this custom dataset was an initial step for this research, which facilitated training and validation of neural network models under conditions that most closely represent actual operating conditions. With a variety of UAV models, distances, and background noise, the dataset is both robust and representative, providing a foundation for meaningful experimentation and resultant accurate acoustic-based UAV detection.

Table 3.1 – Technical parameters of UAV models

UAV Models	**Weight**	**Flying Distance (in meters)**
DJI Mini 2	246 g	2–50
DJI Air 3	720 g	2–50
Qazdrone	150 g	2–50

The technical specifications provided for Table 3.1 were painstakingly sourced from the official DJI webpage [72] for the DJI Mini 2 and DJI Air 3 drones, two of the most popular models of drones used for business and leisure purposes. These technical specifications were instrumental to understanding what each of these drones can do, especially with regards to flight distance and weight, being two of the most critical determinants in understanding how each of these drones performs under various conditions. For the third model of UAV, Qazdrone, the technical specifications were obtained directly from its technical manual, which presented an elaborate discussion of its design and operating characteristics. The flight distances given for Table 3.1 are maximum achievable distances for each model of UAV, as determined by the different manufacturers, and were specifically employed for purposes of this study. It's worthy of mention that while values for Table 3.1 are for generic operating ranges, actual flight distances may be quite different based on several factors, including the conditions of wind, battery life, altitude, and environmental factors. All these variations must be kept in mind when interpreting results and applying them to various practical situations. The total flight distances given by different makers are under ideal conditions, but actual results under field conditions may be different based on technical constraints or environmental issues.

In this research, models were specifically tailored to process sounds in the "WAV" format, which is one of the most popular formats for capturing sounds because it is of high-quality and uncompressed. WAV formatting ensures that original sounds' data is preserved, meaning that the finer points and details of an aircraft's acoustic signature are not lost, which is critical to analyzing sounds in relation to UAV detection. All unmanned aerial vehicle (UAV) recordings during this research were conducted using a high-quality microphone with an auto 16-bit, 44,100-Hz sampling rate, a common configuration for most programs for audio processing. On occasion, for specific technical reasons or to record certain aspects of sounds, an abridged sampling rate of 22,050 Hz was implemented. The target of different sampling rates was determined based upon specific characteristics of the sounds being recorded and how detailed an analysis was desired.

The UAVs were filmed under various flight modes, capturing several different types of motions that these drones usually execute. These include horizontal reciprocating motions, where the UAV traveled back and forth along a defined range, and vertical motions, where the UAV moved upwards and down again in a controlled fashion. The drones were also filmed at various speeds to achieve similar flight conditions to those experienced during actual flight, including low, average, and high-flight speeds. The recordings started once the UAVs were at a short distance to the microphone, to enable their sounds to be recorded at different initial points. The UAVs were also recorded as they moved, with designated initial distances to the microphone and for a range of distances to cover. The different paths and distances for these motions were all controlled and noted during recordings, as presented in Tables 3.2 and 3.3. The intention behind this was to make sure that there were different acoustic profiles included in the dataset to enable a thorough analysis of how the UAV sounds under various operating conditions. The technical characteristics of the recordings made with the "Qazdrone" model were initially supplied in the research [20]. The study provided a basis for further study by providing comprehensive details regarding the flight distances and maneuvers employed during recordings. The distances and maneuvers of the previous study were sustained in this research, ensuring consistency and continuity with regard to experimental conditions. The dataset of the previous research [20] was provided publicly, enabling replication and comparison between different studies conducted within the research domain. With the same dataset and applying the same recording procedures, this research ensures that results are compatible and consistent with previous research, enhancing research with regard to detection and monitoring of UAVs using acoustic data.

Here, repair work noise and an active discussion of a group of many students were just nearby, generating an acoustically dynamic and rich environment. The sounds of construction equipment being used, as well as sounds of conversation and student traffic of a large group, created an additional background noise source that could interfere with the clarity of the UAV's noise. All this was compounded by a mini-truck unloading at an area which was quite close. The mechanical noises of the truck, including those of its engine and goods' movement, added to the prevailing soundscape, and made it ever harder to pick out the acoustic signature of the UAV. All these, including being close to construction and motor vehicle activities, created large variances to acoustic data, and presented an added challenge to measuring the actual acoustic signature of the UAVs under a controlled environment.

The second approach assumed a complete dynamic flight model, with drones traveling at different speeds and directions, mimicking actual operating conditions. This was meant

Table 3.2 – "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)

to replicate conditions under which commercial UAVs are typically used, giving a richer dataset to analyze for the acoustic properties of these drones. More specifically, this dynamic mode of flight allowed for both ascent and descent motions by drones, capturing sounds at various heights. In applying this technique, wide acoustic characteristics were recorded at various heights, but with a specific emphasis on micro-altitudes—such as between one to five meters above ground. These minute changes of altitude were pivotal for analyzing how the sound behaves with changes to the height of the UAV, especially with regards to how environmental elements, including wind or atmospheric conditions, interfere with or affect the drone's sound. With these actual conditions of flight mimicked, the dataset could capture the entire range of noise the UAVs may produce under actual usage, giving a truer picture of what these drones' acoustic profiles look like under different flight conditions.

Table 3.3 – "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

It was through this experiment that I was able to fully describe the acoustic data, as it provided us with a means to differentiate between sounds generated by drones operating in static and dynamic conditions. It is important to recognize that sounds have differing frequency characteristics based upon which dynamics of motion the UAVs are undergoing. Different speeds and heights at which drones are operating generate differing noise profiles, and this introduces heterogeneity into acoustic signatures. As an example, drones operating at higher speeds or at various heights generate more intricate patterns of vibration, which leads to unique characteristics of sounds. It is especially clear with regards to acoustic attributes of drones operating in static positions—i.e., with the UAV not moving—compared to drones under motion, for which dynamic forces acting upon the drone generate additional sources of noise, including motor noise, airflow shocks, and

propeller blade frequencies.
With regard to how important these acoustic variations related to movement are to the overall acoustic profile of UAVs, I specifically designed recording arrangement to record acoustic data under several different conditions. These included static, semi-dynamic, and fully dynamic flight modes. During static mode, I recorded the baseline acoustic profile of the drone, with no external noise related to movement. During semi-dynamic mode, the UAVs were recorded at low-speed flight or during hovering at various altitudes, which is a controlled and steady flight. In full dynamic mode, there were fast-moving drones at various speeds and altitudes, mimicking most realistic flight conditions. With these three different modes, I was able to capture the entire range of acoustic variations that are most representative of actual conditions for flight under real-world conditions. With this wide range of acoustic data, including both the low hum of a still drone and dynamic, high-speed sounds of a drone at full flight, deep learning models for classification of sounds were assured to be trained with the entire range of acoustic features, hence to be accurate at recognizing and classifying various sounds of drones.
In addition, to simulate actual conditions under which UAVs are placed, I partitioned the flight region into various spacial zones, which provided me with different distances to record acoustic data. The spatial zones were created to represent typical distances found during UAV usage. Zone 1, closest to the point that needed protection, ranged from 0 to 15 meters. It mainly represented UAVs at close distances to the microphone, generating sounds with higher intensity and greater low frequencies because of the short range. Zone 2, considered to be the middle zone, ranged between 16 and 30 meters. In this zone, UAVs would generate relatively lesser-intensity sounds as they traveled at longer distances, but still at close enough distances to record major noise profiles. Last, Zone 3, which ranged between 31 and 50 meters, recorded UAVs at far distances, where the intensity of the sounds considerably dropped, and the acoustic data may be affected by environmental conditions like wind, background sounds, and reverberations. All these zones added uniquely to the entire dataset, and thus, there was capability for the models to learn how characteristics of sounds relate to distances of the UAV from the sensor. With this spatial zone methodology, coupled with static and dynamic modes of recording, we compiled an exhaustive dataset that truly reflects realistic UAV operating environments. The segmentation into Zones 1, 2, and 3 accommodated recording of sound data at varying distances, ensuring that deep learning models will be trained to identify and distinguish between the sound characteristics of UAVs at different ranges, from point-blank range to far-range encounters. The incorporation of these well-designed zones in this study design played a critical role in mimicking the actual conditions of UAV operation, where distance and flying conditions dictate what kind of sound signature will be produced by the drone. With this methodology, as depicted in Figure 3.2, I was certain that set of data was ample and comprehensive, and it provided deep learning models with a rich and accurate range of acoustic features for classification.
These main flight zones were, in turn, divided into smaller micro-zones in the course of data recording, with micro-zones being distinguished by a distance of some one meter in between. This methodical segmentation in space was intended to register finer changes in the acoustic signatures created by the UAV at varying distances. This setup not only guarantees a finer mapping of the drone's acoustic footprint, but also paves the way for subsequent investigations that may focus on altitude-based sound distinction and localization. By recording in closely spaced intervals, the dataset gains higher resolution in distance as well as sound variation, thus forming a valuable paradigm for the training

Figure 3.2 – Modeling a Zone-Based Acoustic Sensor System for UAV Distance Prediction

and verification of more sensitive machine learning models that can detect subtle variations in UAV sound profiles. This micro-zoning approach is eventually meant to support increasing the accuracy of UAV distance estimation models in real-world environments.

Table 3.4 – UAV sound dataset.

Name of Classes	Total Duration (s)	Duration of Training Set (s)	Duration of Validation Set (s)
	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

As a result, fully dynamic and semi-dynamic flight mode sound recording techniques effectively recorded a wide range of UAV-specific sounds under different conditions of flight. These techniques provided for a more comprehensive depiction of various auditory features of drones, such as sounds made during horizontal flight, sharp turns, and upward and downward motions. Simulation of various situations with which drones can be involved during actual operation, the recordings yielded rich information regarding various specific acoustic patterns for different drone behaviors.
Additionally, to improve the dataset diversity and make sure that the model was able to distinguish between UAV sounds and other kinds of environmental noise, additional engine-based sounds were deliberately recorded and included in the audio dataset as a sole "No UAV" class. The class was instrumental to capture and study background sounds that might have an impact on detecting UAV sounds. The "No UAV" class contains different kinds of sounds emitted by vehicles, including moving vehicles, small trucks, and motorcycles. The vehicle sounds are usually dominant in urban environments where drones are active, and including them assists in training models to separate the targeted UAV sounds and background traffic.
Additionally, to represent the varied soundscape of an urban area, various construction and repair sounds were recorded and included in the dataset. Noisy construction sounds, including heavy machinery, drilling, and hammering, were loud and dominant during recordings and were included intentionally since these sounds tend to co-occur with UAV sounds under realistic situations. Besides construction sounds, sounds of trains were recorded and included under the "No UAV" category. Sounds of trains passing through are usual for most areas where UAVs may be operating, and inclusion of these sounds still diversifies the database with additional background noise for classifiers to learn to

disregard or filter out. In addition, sounds of human activity were additionally recorded to simulate the impact of public and sporting events within close proximity during the times that the UAVs were being recorded. Examples of these include sounds of individuals moving and communicating, as well as cheering for a football game, to simulate crowds and events, which tend to interfere with acoustic detection by UAVs. Such public and sporting sounds pose an obstacle for models based on acoustics that must isolate sounds from UAVs out of large crowds.
Along with man-made sounds, sounds of wind, rustling of leaves, and chirping of birds have also been included within the dataset. These natural environmental sounds typically exist during UAV flight outdoors, and inclusion of these sounds within the database ensures that the model is able to identify the target sounds of the UAV and non-target sounds naturally occurring around it.
Therefore, the No UAV class of gathered audio data contains an assortment of various types of noise, including car and traffic noise, crowd and public events sounds, natural sounds, and construction and maintenance sounds. The wide variety of these background sounds will make models learned under this data efficient at differentiating between sounds of UAV and various types of confusion sounds. The presence of these varying types of sounds will help in creating models that will be effective classifiers of sounds of UAV and eliminate unnecessary background sounds, thus enabling efficient detection systems.
All of the databases containing class names based upon zones and durations of sounds recorded for all three models of flight are presented here in Table 3.4.

3.2 Model Design

The main goal of this study is to explore the practicability of accurate detection and identification of acoustic signals generated by unmanned aerial vehicles (UAVs) at different distances through various setups of deep neural network models. The motivation for this research comes from increasing significance of accurate drone detection and classification techniques for critical surveillance and monitoring domains like border control, surveillance of infrastructure, and safeguarding no-fly zones. More specifically, this research focuses especially on examining how effectively neural models can be used to identify UAV acoustic signatures to estimate relative distances from a safeguarded point. It is an extremely important task in actual scenarios, where it becomes essential to identify drones at an early stage to have timely response and intervention. Classification of sounds generated by drones based on distance zones can be pivotal for creating layered security measures with a sequence of sensor nodes strategically positioned within spatial regions of interest [73].
In order to accomplish this general goal, several specific research targets were defined. These include: an investigation into several deep learning architectures, including single-layer models for baseline performance benchmarking, stacked multi-layer models for extracting deeper temporal structures, and hybrid models to leverage the merits of several types of neural networks under a single framework. Those models were trained and cross-validated with an acoustic dataset including various acoustic recordings obtained under various spatial regions and flight conditions, incorporating both static and dynamic states of flight to enable a thorough examination.
These earlier investigations, including that by [5], provided proof of concept for ap-

plying , laid out a convincing proof of concept for the application of deep learning architectures—specifically, recurrent neural networks (RNNs) and convolutional neural networks (CNNs)—in the task of UAV sound classification, as documented in 3.5 and 3.6. These studies provided the foundation for determining the advantages of various forms of models when applied to time-varying acoustic signals produced by UAVs. Specifically, earlier works universally substantiated the effectiveness of models based on RNNs, with long short-term memory (LSTM) units being particularly robust in reflecting temporal dependencies and sequential dynamics inherent in UAV acoustic signatures. This advantage is due to the LSTM's capability to remember long-term contextual information, a necessity for processing continually changing sound patterns. In a additional contribution by [74], hybrid architectures fusing CNNs with RNNs also showed substantial improvements against single-model counterparts. These hybrid models successfully combined the spatial feature abstraction capacity of CNNs—valuable for processing spectrogram representations of drone sounds—along with the temporal modeling prowess of RNNs. Consequently, they showed improved performance over a wider variety of drone classes and operational environments. In particular, when faced with environments with excessive background noise, variable ambient environments, or heterogeneous UAV flight characteristics, the hybrid models performed with greater robustness and generalizability. They also performed dramatically well in dealing with interclass variations, effectively discriminating classes between different UAV sound classes even given challenging or noisy acoustic instances. This is indicative of the promising potential of hybrid deep learning models as a potent strategy for overcoming the challenges that characterize real-world UAV audio classification tasks.

Continuing with these initial efforts, this work extends to a broader scope of evaluation with a rigorous comparison of performances across many different architectures. First, simple single-layer recurrent neural networks (denoted as "1L") were considered. While having a relatively simple architecture, these models can be remarkably efficient under situations where computation is simple and inference needs to be fast, as illustrated by Tables 3.6 and 3.7. With an understanding of shallow model limitations, deeper architectures were explored with stacked recurrent neural networks with two stacked layers (denoted as "2L"). The stacked models are likely to extract deeper temporal dependencies and hierarchical sound characteristics, which are needed to perform accurate classification under realistic conditions (Table 3.8).

Apart from these architectures, this study presents a new application of an ensemble of several single-layer neural networks, which produces voting decisions based on the individual model's output. Within this ensemble approach, individual models make contributions to reaching a consensus based on majority voting. Robustness is promoted with this technique through alleviation of bias for individual models and minimization of error causes by noise or unexpected input data distortions. As a result, the ensemble technique proves to be most helpful where there are non-stationary speech signals and interfering backgrounds.

Additionally, hybrid deep learning models (Tables 3.9, 3.10) are explored for development and evaluation. Hybrid models integrate elements of various neural network families to make maximum use of what each family offers. For example, CNNs can perform spatial feature extraction, whereas RNNs have an advantage in temporal dependency modeling. Combining elements into one framework, hybrid models provide enhanced generalization and performance for multi-dimensional acoustic classification problems. The capability of these models to process significantly variable input signals recommends them for use

in UAV detection systems with variable environmental conditions.
During the research, several experiments were performed to evaluate the performance of these architectures under different input conditions and compare results systematically. The primary aim was to find model configurations that, not only produce high accuracy for detection of sounds produced by UAVs at various distances, but are also scalable and adaptive to wider applications. These include range detection systems for drones in real-time, sensor networks for multi-modal surveillance, and edge computing environments for low-latency inference. This study is a contribution to the research area through an extensive comparison of neural network-based methods for acoustic recognition of UAVs. The results provide insightful findings into model selection for deployment cases and point out the capability of employing various learning methods to improve performance. The specific configurations, hyperparameters, and training methods employed for each model design are elaborated upon below for reproducibility and future investigation by others.

3.2.1 Convolutional Neural Network Architectures

In this subsubsection, described how a Deep Learning model architecture is created with CNN neural networks. The task is carried out based on UAV acoustic distance classifications. And this acoustic data of the UAV is input to the model as one-second audio files. Then, the initial layers process this acoustic data through a Melspectrogram layer live. The hyperparameters for Melspectrogram are provided below in Table 3.5. Then, process and feed data through the normalization later to the subsequent 2D CNN layer.

Table 3.5 – Hyperparameter Optimization for the "CNN" in the Proposed Deep Learning Model Series

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

Following normalization, a MaxPooling2D was used straight after the initial 2D Convolutional Neural Network (CNN) to efficiently downsample the input representation and

highlight the most important features of the acoustic data. MaxPooling reduces the spatial dimensions (i.e., time and frequency resolution) of feature maps, which reduces the computational cost of the model and avoids overfitting through imposing a kind of translation invariance. MaxPooling essentially finds the most dominant activations for a region of the feature map, enabling the network to concentrate on strongest and most consequential signal patterns derived from raw UAV acoustic inputs [75].
The initial CNN layer performs a vital function through applying a learnable set of filters to convolve through the Melspectrogram input. The filters function as feature detectors, paying attention to crucial local features like frequency-intensity changes or temporal patterns characteristic of UAV sound patterns. The outputting feature maps provide a learnt input representation emphasizing meaningful features, like motor frequencies or changes resulting from UAV motion and range [76].
In order to expand the learning capability of the model, the CNN layer was repeated again, but with a larger number of filters for deeper feature representations. The second convolutional layer expands what was been learned at low-level features previously, and it extracts more abstract and higher-level audio features through aggregation of low-level features learned at previous layers.
A Flatten layer was subsequently added after the second convolutional layer to convert the 2D arrangement of the feature maps into a 1D vector. The operation is required to ready the data for input into subsequent fully connected layers. The Flatten layer simply unwraps all spatial information contained within the convolutional layers into one extended vector without destroying learned inter-feature dependencies. At last, a Dense (fully connected) layer was included for use at the last stage of model design. It takes care of collating all features learned and coming to a classification conclusion. It operates upon the combined feature vector and computes a probability score for each output class through the use of the softmax activation function. The number of output nodes for the Dense layer is determined by the number of target categories within the UAV sound dataset (i.e., No UAV, UAV for Zone 1, Zone 2, and Zone 3). The Dense layer successfully maps the model's internal representation of the audio features into interpretable results. Each component of this architecture's hyperparameters—ranging from Melspectrogram transformation's sampling rate and window length, to kernel size, activation, and number of filters for CNN layers—were selected based on empirical fine-tuning and domain expertise. These were tuned for high accuracy and robustness under different distances and conditions of drone sounds. A comprehensive overview of these hyperparameters and model components is given below in Table 3.5 [77].

3.2.2 Single-layer Recurrent Neural Network models

This collection of experiments is targeted at the empirical evaluation of deep learning models in a single-layer (1L) recurrent neural network (RNN) architecture, specifically in the context of applying them to classification and processing unmanned aerial vehicle (UAV) acoustic signals. The overall goal was to methodically compare and investigate the performance of various variations in the choice of RNN architecture—SimpleRNN, LSTM, GRU, and BiLSTM—under controlled testing conditions. Input to the models consisted of Mel-spectrograms, computed from raw audio data recorded in flight by a UAV. These Mel-spectrograms provide a time-frequency description of the sound, in essence representing the evolution in time over different frequency bands of the energy in the sound. This representation is especially beneficial in classification problems, as it em-

beds both the spectral and temporal information in the UAV sounds, enabling the models to extract subtle patterns and dynamics that correlate with certain UAV behaviors or distances. By taking advantage of the temporal modeling power in the forms that the RNNs offer, combined with the richness in features in the Mel-spectrogram, this testing phase sought to identify those architectures in the choice of RNN that best capture the temporal dependencies and subtlety in UAV-produced acoustic signals [78].

A number of RNN architectural styles were explored within this study, including SimpleRNN, Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Bidirectional LSTM (BiLSTM). All of these architectures have differing strengths with regards to capturing temporal relationships within the data. SimpleRNN is an elementary base-level recurrent neural network for sequential data processing, whereas LSTM and GRU are advanced versions that can alleviate the vanishing gradient problem, which occurs with long sequences. BiLSTM, by contrast, picks up both forward and backward temporal associations and could potentially lead to a greater understanding of the sequential data by the model.

Kapre library was used for performing data preprocessing for input to the first model layer. Kapre is optimally suited to process audio signals since it supports computation of melspectrograms at efficient rates from raw inputs. This provides high-quality representations to be passed into the model, which simplifies extracting meaningful patterns and performing accurate prediction for the network.

The design of the model is based significantly on results obtained through prior research, where it was found that deep learning architectures with similar features, especially RNN-based layers, perform well at sound classification. As a result, the first Melspectrogram layer was initially set with hyperparameters found in Table 3.6. These include the sampling rate, length of a window, hop length, and number of mel frequency bins, and all these were determined to ensure optimal input features for the model. Following the Melspectrogram layer, data is directed to a LayerNormalization layer. The normalization layer is responsible for ensuring the stability and efficiency of training for the model. Through scaling and shifting of the activations of each feature map, LayerNormalization maintains consistent training data across layers. Normalization through this process assists with minimizing internal covariate shift, where the distribution of activations shifts during training across different layers. Through stable learning, this layer assists with faster convergence and improved model performance overall. This technique takes previous work and applies state-of-the-art deep learning methods to investigate the performance of several RNN architectures for classifying UAV sounds. With a sophisticated use of RNN architectures like LSTM, GRU, and BiLSTM, and with careful adjustment of preprocessing layers and hyperparameters, this experiment strives to push the limits of what can be achieved with sound classification and distance prediction for drones [77].

The input Melspectrogram to the model was transformed with care into the necessary shape for an RNN's processing by including a Reshape Layer (TimeDistributed Reshape). The use of TimeDistributed wrapper allows for this reshape operation to be performed at several different time steps, ensuring that data with time-series characteristics is properly formatted for the sequential operation of RNNs. In this process, the model can process each time step of input data without loss of temporal dependencies inherent within a sequence. Following reshaping, a 128-unit TimeDistributed Dense with "tanh" activation was used. The TimeDistributed Dense layer is needed as it allows for the RNN layer to extract features of desired characteristics out of Melspectrogram data, actually learning sophisticated patterns integral to accurate classification. The "tanh" activation was used

Table 3.6 – Hyperparameter Optimization for the "1L RNNs" Deep Learning Models in the Proposed Experimental Setup

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

for its provision of non-linearity and its capability to learn a wide variety of data distributions.

Following this step, one of the basic recurring building blocks in the architecture was incorporated into the model pipeline, i.e., a version of Recurrent Neural Networks (RNNs)—SimpleRNN, Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), or Bidirectional LSTM (BiLSTM)—was adopted for processing the sequential UAV sound data. These recurrent layers serve as critical for encoding the temporal dynamics native in UAV sound signals, as they are able to learn patterns that evolve in time by making use of both previous inputs and contextual clues stored through the entire sequence. Of these, LSTM and GRU architectures prove ideal for tasks in dealing with long-term dependencies due to the gating mechanisms inbuilt in them, which alleviate the vanishing gradient issue and enable the network to hold important information for longer sequences. BiLSTM, in contrast, extends this feature by being able to process both forward and backward data, in effect allowing the model to leverage past as well as future contexts in tandem, which can prove a boon in performance in dealing with bidirectional temporal patterns within UAV sound sequences. The judicious choice and integration of these RNN versions offer a potent mechanism for modelling intricate temporal associations, necessary for correct deduction of drones' varying sound profiles in dynamic, real-world environments [77].

Following the recurrent layers, the output of the prior TimeDistributed Dense with "tanh" activation was concatenated with those of the recurrent layers by a Concatenate layer.

This type of layer is utilized to concatenate the output of various layers to enable the model to benefit from what both the TimeDistributed Dense and recurrent layers have learned. The concatenation optimizes the framework of the network and improves its capability to identify complex patterns through different levels of abstraction of features. Concatenating these features is a critical step towards enabling the model to distinguish between acoustic signals of UAVs more efficiently.

At this point, subsequent to the RNN process, the model goes through a dense layer with 64 units. The number of units for this dense layer was determined through a line of empirical experiments, which indicated that this provided the most effective performance for the task at hand. The inclusion of this dense layer allows for secondary processing of features that have already been learned prior to forwarding to the last classification layers. In an effort to optimize computation and minimize the risk of overfitting, MaxPooling1D was implemented. MaxPooling1D assists by downsampling the data, shrinking spatial dimensions of the feature map, and only maintaining important information, which improves the performance of the model while creating fewer possibilities of overfitting [79]. It proceeded with passing the data through a Flatten layer, which flattens the multidimensional feature maps into a one-dimensional vector. This is required since fully connected (dense) layers require input data to be flattened. Following flattening, a Dropout layer was used, which randomly sets 50% of the neurons to zero during training. The dropout method is important for preventing overfitting since it disallows the model to become dependent upon particular neurons and compels it to generalize well upon encountering new, unseen data. The model's performance is greatly enhanced under this technique for adapting to new situations. There are two Dense layers involved in the final classification process. There is a first dense layer with 32 units, which is employed for ultimate refinement to enable the model to make its ultimate adjustment prior to output. The second dense layer employs a softmax activation function, which is critical for multi-class classification. The softmax function guarantees that what is yielded is a probability distribution for the four categories, with a score assigned to a specific class as its likelihood. The output yields the model's ultimate classification decision, which classifies the input into one of its acoustic features based on the features produced by analyzing the UAV sounds [80].

In order to improve the performance of the network, Adam as an optimizer was employed. Adam is ideally suited for recurrent neural networks, especially those with adaptive learning rates, as it dynamically adjusts the learning rate during training to facilitate faster convergence and model performance. Furthermore, to suppress overfitting and facilitate generalization of the model, a regularizer coefficient of 0.000001 was selected. This coefficient assists with regulating model complexity and preventing it from fitting too aggressively to training data. The model was trained for 18 epochs, during which the weights were updated, and the accuracy of the model was continually enhanced.

It had already employed this architecture, with success, for detecting and classifying sounds of UAVs, i.e., detecting changes in drone payload or type based upon acoustic signals, in previous research. In the research conducted here, I have attempted to extend this to classify sounds of UAV distances as well. Due to a modification of the make-up of this study's database, actual hyperparameters have been modified, as explained in Table 3.6. The Results section of this paper expounds upon the series of experiments performed with these setups 4.

Following the success of the initial model, the research went ahead to design two more architectures, both retaining the base design of the model outlined above in Table 3.6.

The first solution entailed enhancing the number of cells within an RNN neural network, with all the rest of the layers being as defined above in Table 3.6. The enhancement was done to increase the learning capability of the network to learn sophisticated features and patterns within data by increasing the size of the hidden state of RNN layers.
The second solution entailed reorganizing the network to have two stacked layers of RNN, which allows for deeper temporal dependencies and enhanced understanding of sequence within data.
Table 3.7 presents a comprehensive summary of the updated hyperparameters for the updated neural network architectures, which had been systematically derived and refined from the original parameter list given in Table 3.6. These changes were carefully constructed to increase the networks' capability in dealing with more variable and complex sound input data, like those captured from UAVs in various environmental and distance conditions, without incurring a marked increase in computational overhead. The grounds for adjusting hyperparameters were in the need for improved generalization and resilience of the model, especially when dealing with real-world sound signals that tend to be buried in noisy and unpredictable environments. In this system, LSTM and GRU networks had been selected as they had been enormously validated for modeling temporal dependencies in sequential data, a specialty that is also vital in dealing with continuous sound streams coming from flying drones. These RNN-based architectures had been comprehensively tested in a number of sound classification works and shown to be highly accurate and efficient with a variety of benchmarks [5, 20, 81–96]. Building upon this solid foundation, the present study examines the possible advantage in augmenting the number of neural units in both the LSTM and GRU layers. This improved architectural feature is meant to enable the models to be able to pick up on more subtle patterns in the UAV sound sequences, including fine variations due to drone distance, rotation speed, and environmental noise. Tables 3.7 and 3.8, present the specifications and modifications for each version. These improvements are poised to further extend the limits in UAV sound recognition accuracy and reliability, with valuable improvements in the capacity to identify UAV distance and presence with greater sensitivity and specificity. This strand forms a strong basis for future scalable, real-time drone detection systems in line with deep learning methods with specific focus for acoustic surveillance applications [77].

Table 3.7 – Hyperparameter Optimization for the Experimental Set "1L RNN with 256 Cells" Based on Table 3.6

Layers	**Parameter**	**Range**
Melspectrogram	...	...
...	...	...
GRU/LSTM	Cells	256
...	...	...

Both LSTM and BiLSTM performed quite similarly at recognition tasks, with both having good capability at sequential data operation. Though both performed well, LSTM was noted to have utilized less training time than BiLSTM, hence the choice to concentrate subsequent investigation into LSTM and GRU architectures. The two were specifically tested to see how raising the number of RNN cells within the model affected performance. Experiments indicated that increasing numbers of cells did not have any improvements in recognition, meaning that model performance had reached a state where there was no benefit to be derived from increasing numbers of cells to improve learning capability.

Owing to these results, I did not investigate further increase in the number of cells since results evaluation showed there to be diminishing returns with regards to model performance. I concluded that with existing configurations, the model had indeed been optimized, and additional enhancements to the number of cells were not likely to yield significant improvements. Instead, I decided to investigate various different architectures, and specifically tested the RNN network with a two-layer architecture. This investigation of a stacked, multi-layer RNN network is explored in the section "Two-layer stacked RNN architecture"
My motivation for moving beyond applying additional RNN cells to moving toward a multi-layer network was based upon recognizing that an increase in architecture depth could lead to improved feature extraction and enable the network to recognize temporal patterns of higher complexity. The hope was that this development could make the model process the sequential UAV auditory data more efficiently, and that its results are outlined in the following section. Therefore, with no longer being an option to increase cells within a single-layer network, a natural progression was to look into how increased depth within an RNN could impact model performance for recognizing and classifying sounds of a UAV.

3.2.3 Two-layer stacked RNN architecture

It was built a 2L-stacked RNN-based model by maintaining the design depicted under Table 3.6 and adding two layers with an RNN with 128 cells for each, Table 3.8.

Table 3.8 – Hyperparameter Optimization for the Experimental Setup "2L RNNs" Based on Table 3.6

Layers	**Parameter**	**Range**
...	...	...
GRU/LSTM	Cells	128
GRU/LSTM	Cells	128
...	...	...

The hybrid integration of different RNN architectures, including LSTM, GRU, and BiLSTM, with each other or with Convolutional Neural Networks (CNN), is discussed in the subsequent section. The hybrid model takes advantage of both recurrent networks and convolutional networks to improve the model's capacity to learn both temporal dependency and spatial patterns of data. Whereas RNNs are efficient at analyzing sequential data and modeling temporal relationships, CNNs perform optimally at finding spatial patterns and features. Coupling these two efficacious types of networks, the model utilizes both the sequential learning capability of RNNs and the capability for feature extraction of CNNs, which allows it to process complex, multi-dimensional data with effectiveness. The hybrid model is especially beneficial for those functions, which involve both temporal sequence and spatially dense data, e.g., video and audio analysis, where having a sufficient understanding of both time and spatial features is essential for optimal accuracy. In what follows, deeper insight will be explored into how these hybrid models are implemented and optimized for different purposes, especially with regard to how they enhance model performance for a variety of challenging functions.

3.2.4 Hybrid Models

The last stage of research focused specifically on examining hybrid models that combined different forms of recurrent neural network (RNN) architectures with themselves or combined them with convolutional neural networks (CNNs). Hybridization between these two types of neural networks brings together the strengths of both, creating models that are strong and can process sophisticated procedures. RNNs, which have earned recognition for capturing temporal dependencies, are adept at capturing temporal order and long-range patterns inherent to sequential data. As a result, they are well suited for sequence-driven procedures, where order and longevity are important for good prediction, such as for time-series data or audio signals. CNNs, by contrast, perform spatial features and various structures within data well and are well suited for procedures that process spatially dispersed data, like image or acoustic signal analysis. When used together, CNNs and RNNs complement each other by harnessing the CNN's spatial feature efficiency with the RNN's temporal dependency understanding. With this hybridization, both temporal and spatial aspects of data can be considered at once, leading to improved performance, for example, for detection of UAV acoustic signals.
Examination of hybrid models was meant to bring about new avenues of opportunity in the field of UAV audio signal detection. The methodology harnessed the strengths of CNNs and RNNs to improve the model's capability to detect and classify audio signals, both capturing structural patterns and dynamic temporal dynamics inherent in the audio data. Merging these two forms of neural networks, research attempted to design a richer model that could tackle the challenges of UAV-sound classification. First, as illustrated in Table 3.9, two various RNN types were tested together to see how different recurrent structures could cooperate together. The initial test was conducted to see how different RNN models could be combined to make the hybrid model perform better in both capturing short- and long-duration dependencies within the audio content. The various combinations were tested for how well they can process and comprehend the sequence of the audio signals, with an emphasis on selecting the most effective configurations. The outcome of this experiment laid a basis for subsequent optimization of hybrid models, paving the way for more sophisticated methods in UAV audiosignal detection [78].

Table 3.9 – Hyperparameter Optimization for the Experimental Setup "LSTM-GRU" Based on Table 3.6

Layers	**Parameter**	**Range**
...	...	...
LSTM	Cells	128
GRU	Cells	128
...	...	...

Following exploratory experiments were carried out employing a combination of CNN and RNN types of networks, Table 3.10.
In this research direction, It focused on investigating hybrid models, focusing specifically on blending Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) through different combinations. The motivation for this direction was to make use of the respective strengths of both architectures and improve the performance of the system for UAV (Unmanned Aerial Vehicle) sound prediction.
Convolutional Neural Networks (CNNs) excel at learning hierarchies in the spatial do-

Table 3.10 – Hyperparameter Tuning for Proposed conv2D-LSTM and conv2D-BiLSTM Architectures

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

main and parsing structured attributes from raw input data like spectrograms, a two-dimensional time-frequency representation commonly used for representing sound signals. Their architecture, consisting of convolutional layers, pooling, and nonlinear activations, facilitates CNNs to automatically identify local patterns like harmonics, modulations in frequency, and other sound signatures characteristic of distinct sound events or classes. This property renders CNNs a valuable asset in tasks involving classifying sound with the help of spectrogram input from the sound generated by UAVs, wherein discriminating, intricate features may occur at differing frequencies and time periods [97]. In contrast, Recurrent Neural Networks (RNNs), specifically their advanced forms like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are geared towards representing temporal sequences and dependencies and can capture long-term dependencies over time. These models store internal memory states, and they can remember and utilize information from previous time steps in order to make predictions for subsequent time points. This feature for representing temporally is particularly important in sound signal analysis, where the time-series aspect of sound—e.g., changes in drone motor pitches, or repeating noise patterns—plays a vital role in proper identification and classification. Therefore, the tasks involving recognizing sound dynamics in time scale well with the application of RNNs, and they balance the performance capabilities of CNNs for being effectively combined in hybrid architectures with benefits from both space and

time feature representation [98, 99].
It was tested various hybrid model architectures that integrated CNNs and RNNs in this study. The purpose was to employ CNN layers to extract salient features from the acoustic data and subsequently input these features into an RNN layer, which subsequently modeled temporal dynamics of the sequence of sounds. It was sought to integrate CNNs' capability to extract features and RNN's capability to process sequence to develop models that could predict sounds of UAV under different conditions with greater effectiveness.
These hybrid frameworks were tested and analyzed based on various measures, namely, accuracy, precision, recall, and F1 score, to assess how well and under what situations they performed. The experiments were conducted to find out the best configurations for CNN and RNN hybrids that yielded the most accurate and stable predictions for detecting UAV noise, accounting for differing patterns of flight of the UAV, environmental noise, and other extraneous conditions. In these evaluations, I attempted to show how hybrid CNN-RNN models can provide meaningful enhancements to single-model methods and enable future advanced and trustworthy UAV sound prediction systems. The ability to effectively integrate CNNs and RNNs into these models can potentially enable improved monitoring and detection of UAVs, which is critical to numerous fields including airspace management, security, and environmental monitoring.

3.2.5 The Voting system

Prediction of UAV recognition via weighted voting based on multi-trained single-layer RNN models is an additional critical research direction within this research study. The research methodology was chosen with great care to tackle the problem of attaining high recognition rates through a combination of several expertise-intensive models' predictive powers as opposed to one model structure alone. Weighted voting's underlying basis is its capacity to tap into several model viewpoints while assigning to every model an influence based on its performance and dependability [100].
Under a classical majority voting system, every model votes an equal number toward a prediction conclusion, irrespective of performance under certain types of data or experimental conditions. While ensuring ease, this system does not consider differences between model strengths, yielding non-optimal prediction results, especially under complicated conditions, as in the recognition of UAV sounds [100, 101]. Conversely, a weighted voting system presents an added dimension of adaptability through different weights assigned to individual models participating in voting. These weights depend on the trust and performance values obtained during initial tests or validation phases [100, 102].
Weight assignment within experimental model underpinned by performance-based assessment. More specifically, models that regularly achieved high accuracy, precision, recall, and F1-score for UAV audio data were weighted more heavily. As an example, a GRU model which performed better than other RNN variants for short-range UAV recordings received increased significance within the ultimate decision-making stage. On the other hand, models that showed relatively poorer performance received weaker weights, so as to proportionally minimize their impact upon ultimate prediction results. Overall, not only did this performance-sensitive methodology make system more robust, but it provided for more dynamic and equitable incorporation of several models, each with specific strengths across different data subsets [101].
At the decision point, every model in the system cast a vote for a given class—e.g., "UAV," "No UAV," or a specific UAV type—but as opposed to unweighted voting con-

ventions, this research adopted a weighted voting approach where a vote from each model was weighed by a precalculated weight that represented how accurate or reliable it was, as ascertained through validation. A weighted sum from every class was determined, and the class with the highest total weighted value was chosen as the overall prediction. This system guaranteed that models with higher performance influenced the decision far more, with less accurate or inconsistent models proportionally having less effect, ensuring robustness and precision in the classification decision. The weighted voting system resulted in a consensus derived from the best-performing models rather than a majority, eliminating the possibility for mediocre or biased answers to control the outcome. This is in line with the objectives of ensemble learning to prevent systems from being lessened by poorly performing units and is especially beneficial in mission-critical applications like UAV detection, where precision under diverse environmental conditions is paramount [100].

Effectively, the use of a weighted voting approach gave a sophisticated means of combining the results of several RNN models with a view to ensuring maximum strength and offsetting weakness. The ensemble methodology was especially useful in managing the variety and range found with actual UAV recordings, including how environmental noise, range, and type of drone can affect prediction performance [77].

Experiments conducted with research involved mainly carrying out an extensive study of different deep neural network architectures used for recognition of sounds in UAVs and enhancing their functional abilities. Another goal of the study was to find out which model is most efficient by analyzing and contrasting recognition benefits provided by various models. All hybrid models' performance was investigated systematically and experimental results were contrasted with those of the prior models.

The wider relevance of this experimental approach is that it contributes to increasing literature on both neural architecture combination and ensemble learning. In demonstrating that weighted voting between different RNN models with strategic integration is capable of achieving superior recognition performance, this research presents a practical roadmap for future studies not only for UAV detection systems but for all types of sound-based classification as well.

Furthermore, this study provided insightful contributions to the efficiency of hybridization between different neural network structures, particularly with regards to acoustic pattern recognition. Along with analyzing individual types of RNNs like GRU, LSTM, and BiLSTM, the research also considered hybrid structures with interconnections, where these RNNs were combined with CNNs to build deep hybrid structures. These CNN-RNN hybrids combined CNNs' capability for feature extraction with RNNs' ability to model sequential patterns, and thus were well suited for handling complex, time-varying acoustic signals like those generated by flying UAVs.

Although the "Results" section of this paper offers a comprehensive report summarizing different findings—such as the individual weights given to each RNN model in the voting process, in-depth hybrid vs. single-model architecture comparisons, and visual representations of classification accuracy for multiple different experimental configurations—their collective outcome plays a greater role than mere verification. They highlight the practical utility of the proposed methods and provide a solid basis for the future evolution of scalable, real-time identification systems for use in unmanned aerial vehicle (UAV) detection tasks fueled by deep learning. One of the key take-home points is the utility gained by the combination of weighted voting and hybrid architecture designs, which created dramatic improvements by both detecting temporal and hierarchical structure in UAV acoustic signals. These architectural designs not only remain technically pertinent, but

they also hold a strategic role in building next-generation audio understanding capabilities that can operate soundly in disparate, changeable environmental conditions. The knowledge obtained through this investigation reaffirms the great promise that lies in ensemble methods with respect to deep learning and emphasizes the utility of architectural testing and optimization. Ultimately, this helps to move the capabilities of the machine learning systems forward toward being deployed in practical, high-stakes applications like security surveillance, military surveillance, and unmanned airspace management tasks involving UAVs [76, 77, 101, 103].

Chapter 4

Results

The purpose of this research is to examine and assess the viability of advanced deep learning methods for estimating unmanned aerial vehicles (UAVs) distance based exclusively on their generated audio signals. Over recent years, acoustic-based methods for detecting and localizing unmanned aerial vehicles have attracted serious consideration because of their non-invasive, low-cost, and passive features. Since, during flight, UAVs produce characteristic sounds, mainly through propellers and motors, these sounds can be an effective solution to estimate proximity or distance. As such, this research section is exclusively devoted to a detailed investigation of deep learning architectures, i.e., recurrent neural networks (RNNs), convolutional neural networks (CNNs), hybrid model structures with both CNN and RNN merits, and ensemble learning methods employing weighted voting frameworks [48].

In order to provide strong model training and fair estimation, the raw acoustic data obtained during experiments was partitioned into three clear subsets: a training subset of data for model learning, a test subset of data to examine generalization performance, and a validation subset of totally unseen speech recordings to test adaptability of the model to new data. The partition was done to prevent overfitting and to examine the predictive performance of models under realistic circumstances.

The methodology used to undertake this work is based on previous research within the field, most significantly that undertaken in [5], which tested and proved effective the use of deep learning methods for detecting drones based upon their acoustic profiles. Expanding upon the structural investigations and results of said previous work, this research advances deep learning's usage by moving beyond simple drone detection to that of estimating distances of UAVs. A key innovation of this research is based on the integration of ensemble learning through weighted voting schemes. In contrast to simple majority voting, where all involved models have an equal say, weighted voting allows for variable influence for different participating models, based on specified performance measures, for example, validation accuracy or task-dependent recognition ability. With this, it is feasible to have a more intelligent, fine-grained decision process, with stronger models exerting more influence in the prediction. Additionally, hybrid models based on CNN layers, which are good at capturing localized spectral features from representations of sounds as Mel spectrograms, and RNN layers, which are good at capturing temporal patterns across time, are systematically built in this study. Hybrid configurations based on these model types are meant to make model robustness and accuracy for unmanned aerial vehicles distance prediction better by exploiting both models' available strengths.

In summary, this chapter introduces an extensive review of deep learning-based methods for acoustic distance estimation for UAVs, including individual model structures and ensembling methods. Based on systematic experiments and assessments, the research attempts to find out the most efficient model structures and methods to enhance the capability of UAV detection under actual-world audio conditions. The utilized models, training methods, and testing methods are explained comprehensively below.

A series of controlled experiments were performed with a standard personal computing setup, i.e., a laptop with a processor being an Intel(R) Core(TM) i5-8265U processor at a base frequency of 1.60 GHz. All model creation and experimental steps were carried out by employing the Python programming language within the Spyder Integrated Development Environment (IDE), which was found to have an intuitive user interface and efficient script running, debugging, and visualization facilities. In order to enable efficient manipulation of audio data, user-friendly Python libraries were installed, which included the Kapre library—intended for on-GPU based audio preprocessing—and various necessary deep learning layer libraries needed for model creation, training, and evaluation. These libraries were crucial for implementing Mel-spectrogram-based audio feature extraction and facilitated integration of sophisticated neural network architectures.

At the initial stage of experimentation, several deep learning models were designed and successively trained under different numbers of epochs to study convergence behavior and generalization performance. Some initial exploratory training was performed with epoch values between 25 and 50 to measure at which point models realized stable learning curves without developing signs of overfitting. Following this study, a standard benchmark of 18 training epochs was chosen for all experiments based upon two considerations. First, all models converged and realized a stable "good fit" at around 18 epochs, which indicated that training for longer past this point realized diminishing returns or, alternatively, overfitting for a few models. Second, consideration of a constant number of epochs ensured fair and unbiased comparison between different model architectures. All models being trained under consistent conditions, especially regarding training time, eliminated variability based upon different learning curves or overtraining and enhanced the validity and interpretability of the comparison results.

Once training for all models was complete, trained weights and architecture configurations were saved with Hierarchical Data Format version 5 (HDF5), with an extension ".h5" being used for these files. The file type is well-supported across the deep learning community and is well-suited for saving Keras and TensorFlow models, and its usage allows restoration and reusability with no effort whatsoever for future prediction purposes. The saving of trained models not only ensures reproducibility of results and supports future testing, but it creates a framework for deploying those models into actual application environments, including for ensemble-style environments like weighted voting. In maintaining these trained models in an accessible and formalized fashion, the research allows for reusability and practical applicability, both requirements for moving forward with development for UAV detection systems and integration into operational streams.

As described in previous segments, the original dataset was partitioned into three different subsets with care to facilitate the strong development, testing, and validation of the suggested models. The total length of all recorded audio data equaled 16,962 seconds, representing a large and varied dataset for acoustic distance estimation. For ensuring the integrity and independence of each stage of evaluation, the test and validation datasets were sourced and segregated from training set before model training started. A dedicated validation set comprising around 10% of total recordings from all classes was reserved.

The validation files were kept entirely out of training and test procedures and were utilized solely for validation alone to evaluate models' generalization performance based solely on unseen data. The major intention of this validation process was to find out what errors could be there, measuring stability of output provided by models, and building confidence for operational environments. Secondly, for enabling recognition requirements in real-time, all the audio pieces were preprocessed to make them all have a constant length of exactly 1 second. Standardization was imperative to facilitate time-driven prediction and equalizing input dimensions for all neural network models. After isolating the validation set, the remaining 90% (corresponding to 15,266 individual audio files) became the total dataset for testing and training. The subset was later partitioned by a ratio of 80:20, with 80% being reserved for training and 20% for the test set. Such partitioning ensured that models got to see a varied and representative training set and provided an ample and independent test set for post-training performance estimation. Such a management of data was pivotal for ensuring serious evaluation of models' accuracy, effectiveness, and suitability for real-time prediction of UAV distances. With regard to feature extraction and transformation of the audio signal, this research benefits from building upon prior research, and specifically, upon the research provided in [5], which thoroughly examined preprocessing drone-audio data. Drawing upon the conclusions reached within it, the Mel-spectrogram transformation layer was adopted as the main audio preprocess step for this research as well. Mel-spectrograms, which offer a time-frequency representation of sound consistent with that perceived by the human auditory system, were used to transform raw audio signals to 2D representations for convolutional processing. The effective use of this process in prior research coupled with its effectiveness at capturing salient acoustic features for a variety of different UAV types and distances made it an attractive inclusion for use within the deep learning models of this research. The use of this sound process layer reflects both continuity and advancement of research within this subject area and a commitment to drawing upon well-established best practice within audio-derived UAV distance estimation.

4.1 Single-layer RNN and CNN architectures

The second subsection is reserved for examining and analyzing lightweight deep learning models, which are well-suited for situations where computational resources are limited or for environments where there's a demand for real-time performance. The primary goal of this stage was to assess the base model structures prior to moving forward to developing more advanced or computationally expensive architectures. Lightweight models are essential for embedded systems or edge devices, like unmanned aerial vehicles (UAVs), where there's limited power and limited capacity for processing. In this regard, the research sought to find architectures that can provide an ideal trade-off between model accuracy and computational cost.

Experimental development started with the deployment and experiments with convolutional neural networks (CNNs), which are renowned for extracting meaningful spatial features from input data. The CNNs were used as a baseline for performance during initial phases of research. Different architecture parameters were tweaked, including numbers of convolutional layers, filter size, and inclusion of pooling layers, to find out what configuration was most efficient to process UAV audio signals. These models were kept relatively shallow to maintain low computational cost, but deep enough to be insightful and pick

out patterns within the data.
After CNN-based experiments, single-layer recurrent neural networks (RNNs)-based models, including SimpleRNN, LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), and BiLSTM (Bidirectional LSTM), were considered. These structures were chosen for their proven capabilities for capturing temporal dependence and sequential patterns of audio data. Because most UAV sound signals involve minute changes with respect to time, models based on RNNs were found to be especially helpful for monitoring these changes and improving recognition performance. All the variants of RNN were tested separately in a lightweight mode — i.e., with just one recurrent layer — to preserve computational convenience and evaluate performance both for accuracy and training duration. Each model's performance was tested through the recognition accuracy values obtained through training and validation processes. The recognition accuracy values were instrumental in understanding how rapidly and efficiently each model learned to recognize UAV audio signals. As you can see through Figure 4.1 [77], models were compared and contrasted with each other to compare their capability and robustness during training as well as how well each model generalized. With these results, it was determined which of the lightweight models worked optimally with the dataset at hand and formed a precursor to creating improved or hybrid models for future development in subsequent parts of this research.
Overall, this first stage of CNN and single-layer RNN-based architecture testing was essential to baseline performance measures, to recognize potential temporal or spatial modeling strengths, and to make sure that models produced were accurate and efficient in resource usage. These results were instrumental to informing architectural decisions made later in the study.

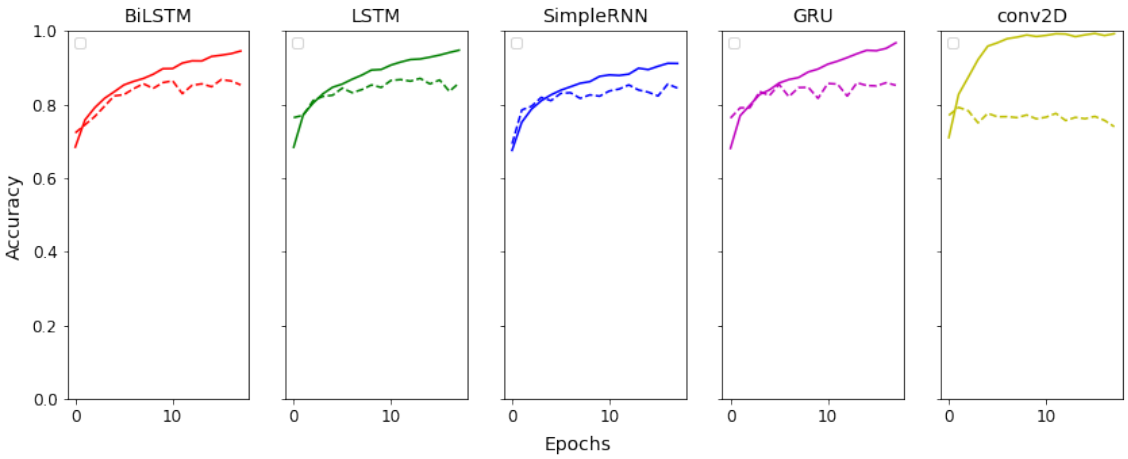


Figure 4.1 – Confusion Matrix of CNN-Based Deep Learning Model

For all of these deep learning models, which were established and experimented with during this research stage, a detailed performance analysis was carried out to identify the strengths and weaknesses of the respective architectures. In this detailed analysis, there were various key measures and visualization tools used to evaluate how well the different models could recognize the audio signals related to UAVs. The major emphasis was laid on recognition accuracy, which was used as a reference to compare model performance. But for a better understanding of prediction behavior, additional methods of evaluation, which included generating and analyzing recognition scores and confusion matrices, were included for the study.

Recognition accuracy was monitored for both training and validation sets for every model throughout several epochs. These recognition curves were instrumental in giving insight into how well the model was learning, whether it was underfit or overfit, and how consistent its performance was while training. With these trends, decisions were made for hyperparameter adjustment, model architecture modifications, and training approaches. As an example, if a model's training accuracy was high, yet its validation accuracy was much worse, it was an indication of overfitting, which would lead to changes like the application of regularization or dropout layers.

Along with accuracy curves, confusion matrices were employed for gaining a clearer picture of classification performance. The confusion matrix is a generally accepted measure for classification problems, providing a granular view of model performance for each class individually through an aggregation of correct and incorrect classifications. In part, confusion matrices were created using the validation set and shown across each model. The matrices showed actual vs. predicted labels and were sized in units of number of audio files, thus providing a concrete understanding of classification performance. The confusion matrix for the CNN model is presented in Figure 4.2, demonstrating how well the CNN performed at separating different categories of UAV sounds. Analogously, confusion matrices for different RNN model types—including SimpleRNN, GRU, LSTM, and BiLSTM—are presented in Figure 4.3, providing comparison insight to their temporal classification capacities. In addition to visual analysis, quantitative performance results were also compiled and summarized within tables. Table 4.1 shows overall model performance as percentages, based on the respective confusion matrices. The table provides a model-by-model comparison based on classification performance, thus illustrating which architectures performed best with this study's constraints and conditions of data. The inclusion of relative (percentage) and absolute (counts as presented by the confusion matrices) measures of classification accuracies ensures performance can be interpreted both relatively and absolutely.

As a comprehensive, multi-faceted process including accuracy charts, confusion matrices, and summary presentations via percentage based upon classification, it allowed for a thorough, rigorous assessment of all lightweight architectures as a group. It not only illustrated CNN and RNN classification performance but provided direction to improve and develop still newer, more sophisticated structures examined during the later phases of investigation.

Experimental evaluation showed that CNN-based networks exhibited poor reliability to estimate UAV distances, especially under range conditions with longer ranges. Its performance not only suffered with regard to recognition accuracy at longer distances but also showed inconsistencies within accuracy curves, reflecting instability of learning and generalization. These flaws underscored the weakness of spatial-feature-based models to perform well with image-like data that is heavily dependent on temporal dynamics, like range prediction for UAVs. On the contrary, recurrent neural network (RNN) structures—SimpleRNN, GRU, and LSTM—showed relatively stable and accurate recognition performances under the same group of experiments. Their ability to capture temporal dependency made them well suited for sequential data like acoustic signals, where temporal variations have an important contribution to classification.

Even with relatively enhanced performance being achieved by RNN models, none of the separate networks reached an adequate degree of recognition reliability needed for accurate UAV range prediction under actual conditions. Due to this shortcoming, the research went forward to investigate ensemble methods—concretely, a voting technique—

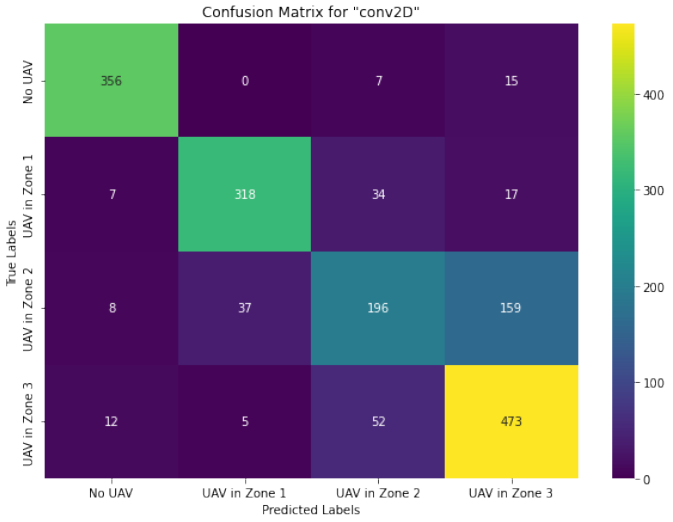
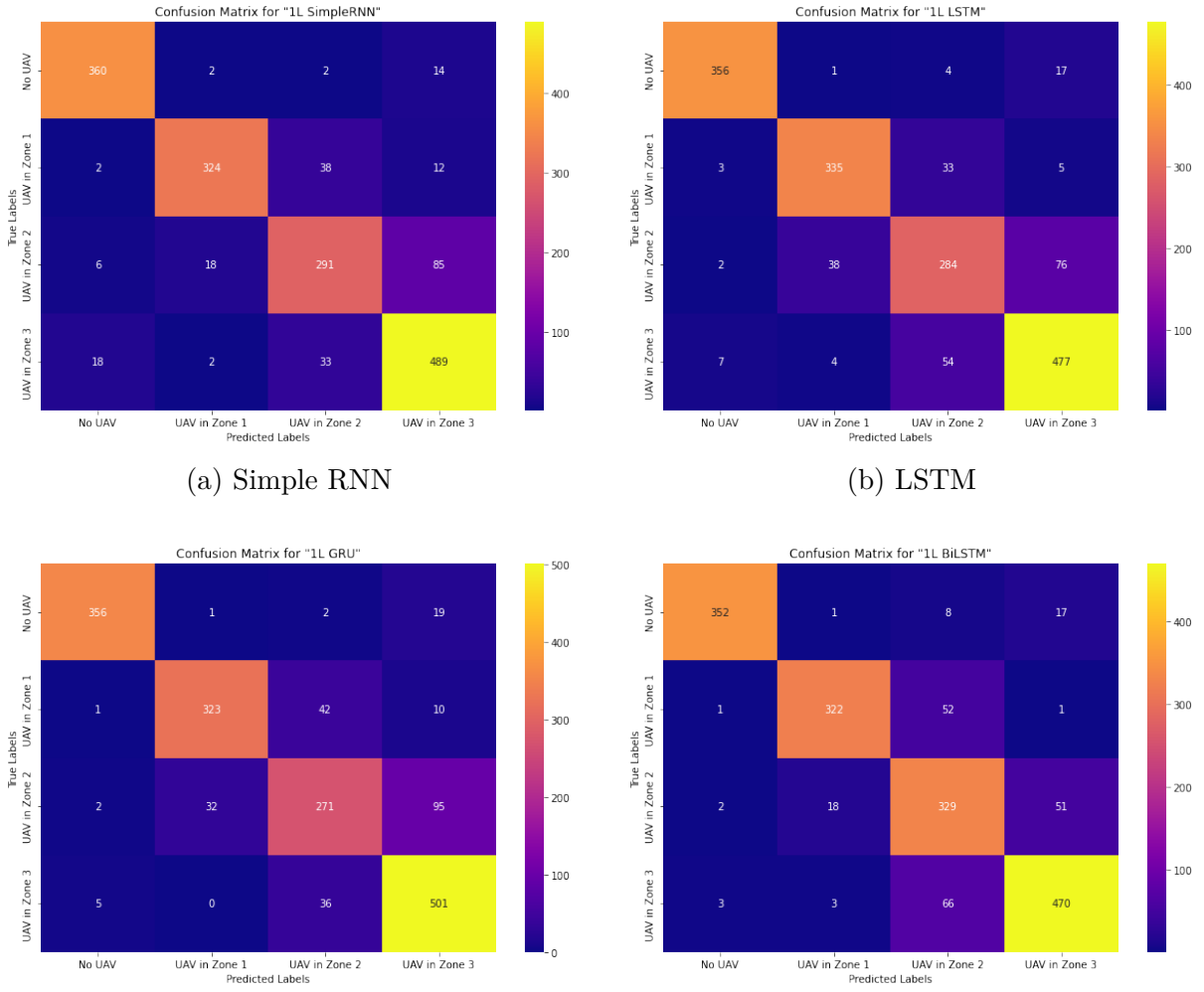


Figure 4.2 – Confusion Matrix of CNN-Based Deep Learning Models

as a solution to increase recognition reliability through a group of classifiers' summation of strengths. The ensemble approach sought to utilize different models' shared strengths to increase collective classification accuracy as well as robustness.

In order to thoroughly test the performance of the models that were created at the onset of the study, five independent trial experiments were carried out. The experiments involved measuring the predictive accuracy of the networks through the use of established performance measures: Precision, Recall, and F1-score. These measures provide an unbiased approach to measuring performance classification, especially for situations where class distributions can be unbalanced or where both false negatives and false positives have serious penalties. Classification reports were created and examined for each model to identify how well the models classified and differentiated between different classes of UAV signals. These evaluation measures were computed for these experiments as per the methodology outlined in [6], ensuring reproducibility and validity of results. The confusion matrices, which show both correct and incorrect classifications of audio files for all classes, were originally displayed as absolute numbers, as indicated in Figure 4.2. It provided a concrete reference for raw results of classification. Supporting this, Table 4.1 displays these confusion matrices as a percentage, which offers a normalized and comparison basis to observe model performance for different classes and experiments. Table 4.1 contains detailed statistics for Precision, Recall, and F1-score as well, which offers an extensive quantitative overview to better comprehend both strengths and needs for improvement for each model. This extensive assessment paved the ground for subsequent research steps, which constituted hybrid and ensemble learning architectures for improved predictive capacity.

Overall, all three networks, LSTM, BiLSTM, and GRU, provided improved performance over that of CNN for all experimental conditions. The LSTM and BiLSTM networks, specifically, provided improved recognition of objects, including with background noise. In spite of that, GRU showed an impressive performance for object recognition at far distances with a greater degree of consistency than LSTM and BiLSTM. Conversely, the BiLSTM model performed with improved accuracy and resilience for calculating drone distances, most specifically in noisy conditions. But for object recognition at far distances, GRU performed better than BiLSTM.



(a) Simple RNN

(b) LSTM

(c) GRU

(d) BiLSTM

Figure 4.3 – Confusion Matrices of Deep Learning Models Based on 1L Recurrent Neural Networks

With these results, it was evident that no one model could outperform others for all conditions consistently. The subsequent stage of research will henceforth involve investigating whether there is benefit to be gained through an integration of the strengths of GRU, LSTM, and BiLSTM networks. This will be provided through a voting system, where all the abilities of these models can be amalgamated towards increasing overall classification accuracy. Through an integration of the abilities of all three networks, this hybrid model will be improved for various hard-case scenarios, including object identification at short and long distances and in noisy conditions. The subsequent section will describe implementing and implementing this voting system based ensemble model.

4.2 Prediction through Weighted Voting with 1L RNNs

In this section, an attempt is made to integrate the strengths of the stronger neural networks investigated within the last section by means of voting. A voting scheme, which compiles prediction made by several models into one output, has been well-established as a good methodology to improve machine learning model performance. Using this tech-

Table 4.1 – Evaluation of Model Performance and Predictive 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

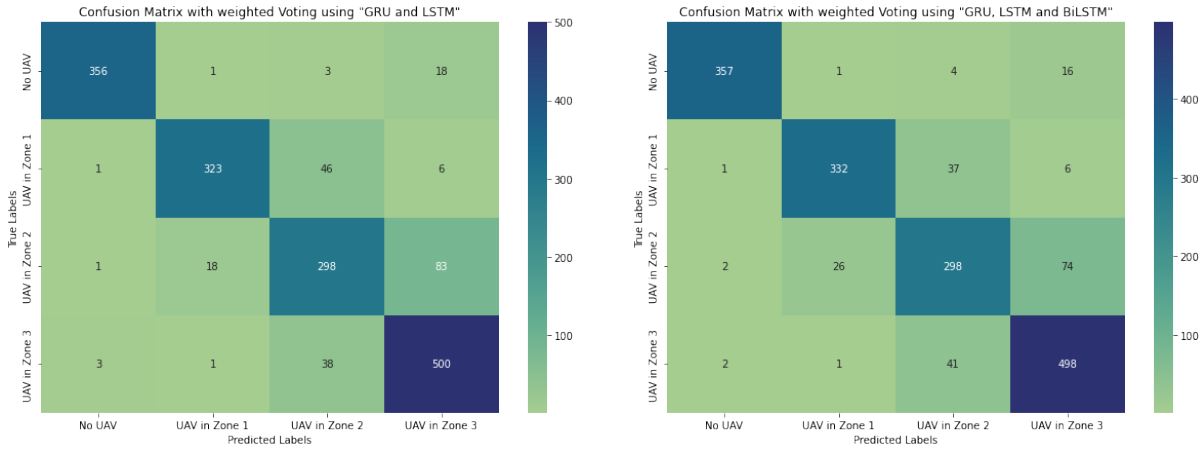
nique, the study tries to preserve the merits of each model and eliminate its specific weak points, which eventually leads to an overall improved classification performance. The major concept of this methodology is to compare various networks' prediction and make a collective decision for generating a reasonable and accurate model.

A Weighted Voting System, an advanced variation of the simple voting mechanism, was used by the study to integrate the neural networks' forecasts. Various models received different weights based on how well they performed individually, enabling the system to emphasize better-performing models at decision-making. The study tested two major combinations of neural network models:

Combining GRU and LSTM Networks – The GRU and LSTM networks, with proven recognition performance under various environments, were combined to make use of both of these networks' capabilities. The weighted voting scheme for this integration is presented in Figure 4.4a, which aggregates both GRU and LSTM networks' prediction to improve object recognition performance.

Combining GRU, LSTM, and BiLSTM Networks – Within this second solution, the prediction of all three networks—GRU, LSTM, and BiLSTM—is combined via weighted voting. This blend attempts to leverage the unique strength of each network, i.e., BiLSTM's bidirectional dependency capability, and holds on to the consistency of GRU and LSTM networks. The weighted voting technique for this blend is depicted under Figure 4.4b, where the output of the three networks is compared and merged to produce an improved prediction. By incorporating these models using weighted voting, this re-

search expects to increase the overall performance of UAV audio signal detection with emphasis on improved classification and strong recognition under different environmental conditions.



(a) Voting System GRU-LSTM (b) Voting System GRU-LSTM-BiLSTM

Figure 4.4 – Confusion Matrices of 1L RNN Voting Models

Table 4.2 – Evaluation of Model Performance and Predictive 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

In research, a weighted voting system was employed to rank individual models based on their respective performance toward recognizing UAV audio signals. In particular, for the first model combination—GRU and LSTM—priority weights were set as follows: The GRU model (M1) was weighted at 0.4, and the LSTM model (M2) was weighted at 0.6. These weights represent the relative performance of each model toward recognition purposes. The LSTM model, with its higher capability to learn temporal dependencies and long-range association between the audio signals, received a higher weighting to increase its impact toward the decision process. The GRU model, albeit efficient, received a relatively smaller weight since its capability toward object recognition at greater distances was considered. The combination of these two models was expected to benefit both methods, as indicated by Figure 4.4 (a).
For the second combination, where all three GRU, LSTM, and BiLSTM models were combined, the weights were assigned with priority as follows: M1 (GRU model) was assigned a priority of 0.35, M2 (LSTM model) an assignment of 0.45, and M3 (BiLSTM model) an assignment of 0.2. The weights were selected with reference to the strengths of the three models noticed in different aspects of recognition. The LSTM model, which performed well for sequential patterns and temporal dependencies, was assigned the max-

imum priority of 0.45. The GRU model, which performed well for long-distance recognition, was assigned a priority of 0.35. The BiLSTM model, which was good for capturing bidirectional dependencies, was assigned a priority of 0.2, representing relatively fewer contributions than the first two models. Combining these three models, giving their different strengths, was meant to deliver a comprehensive and robust recognition performance. The weighted voting system's flexibility is that it can fine-tune the contribution of each model to be included in the ultimate decision-making process. The weights can be adjusted according to the recognition capacities of individual models. Dynamic adjustment to the data and task characteristics is thus ensured, with most accurate models contributing to the ultimate output. In this instance, upon reviewing both sets of results, the second methodology, which merged GRU, LSTM, and BiLSTM models, had much better recognition performance.
The effectiveness of this hybrid solution is actually proven through its performance based on two key performance indicators: Figure 4.4 and Table 4.2. The weighted voting framework, with contributions of each model based on its performance, greatly improved accuracy and dependability of UAV range recognition. The process enabled the models to cooperate synergistically, leveraging individual strengths to benefit the system's overall performance, enhancing it based on different environmental conditions. The performance through these indicators established that GRU, LSTM, and BiLSTM networks' combination performed best, yielding a viable solution for both UAV audio signal detection and range prediction.

4.3 Single-layer RNN Architectures with More Cells

As noted in Subsection 4.1, GRU, LSTM, and BiLSTM networks performed best at recognition, outclassing other architectures tested, especially for application to UAV audio signal classification. Their temporal representation capability and sequential data management properties made them promising targets for deeper investigation. Due to their promising performance, an extra set of experiments was carried out, specifically with an emphasis on GRU and LSTM networks and increased numbers of neural units, or memory cells, to determine whether increasing their capacity may improve model performance. The intent was to investigate how model complexity affects recognition accuracy and to see if enhancing representational power within the network would produce meaningful improvements. Through gradual variation of the number of units within recurrent layers, I sought to test for GRU and LSTM model scalability and learnability upon being presented with increased internal complexity structures. This step was instrumental to uncovering optimal configurations for high-performance recognition systems for the application to audio based UAV detection. In this line of investigation, the BiLSTM model was also considered for additional testing with a greater number of neural units, under the assumption that its performance indicated noticeable improvements. The basis for this choice was to investigate whether bidirectional processing ability of the BiLSTM benefited to a greater extent with larger capacity than its unidirectional counterparts. Nevertheless, as evident with Table 4.3, experimental results indicated no appreciable enhancement of recognition accuracy and robustness of the BiLSTM model with a rise in number of units. Gains in performance were negligible and did not justify additional computational cost and complexity. The investigation thus did not continue with carrying forward this line of testing for the BiLSTM model. Instead, emphasis was kept with

Table 4.3 – Evaluation of Model Performance and Predictive 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

the GRU and LSTM networks, which performed relatively more promisingly under comparable modifications. The choice served to streamline investigation by focusing effort on models with a greater potential to deliver practical improvements.

4.4 Deep Learning Models with 2L Stacked RNNs

In this subsection, the GRU and LSTM neural network architectures, which were highlighted in the previous section for their strong recognition capabilities, are further explored through a different architectural approach—specifically, the stacking method. Stacking involves layering multiple recurrent units on top of each other to form deeper network structures, allowing the model to learn more abstract and complex temporal patterns from the input data. This technique aims to enhance the representational capacity of the networks by enabling the upper layers to capture higher-level temporal features that are built upon the lower-level representations. By applying the stacking method to the GRU and LSTM models, this study seeks to evaluate whether increasing the network depth contributes to improved accuracy and robustness in UAV audio signal classification tasks.

Table 4.4 – Evaluation of Model Performance and Predictive Metrics

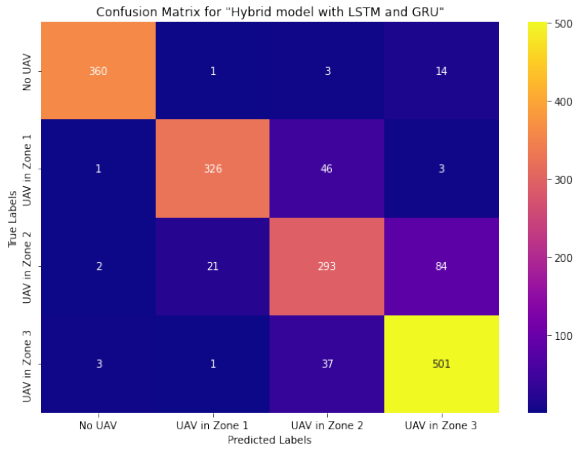
Model	Classes	Precision, %	Recall, %	F1-Score, %
"2L GRU" with 64 cells				
	No UAV	99	92	95
	UAV in Zone 1	92	87	90
	UAV in Zone 2	76	76	76
	UAV in Zone 3	83	90	86
"2L LSTM" with 64 cells				
	No UAV	95	**96**	96
	UAV in Zone 1	91	**89**	90
	UAV in Zone 2	81	**76**	78
	UAV in Zone 3	86	**90**	88

Experimental results during this stage of the study showed that it was more effective to use a stacked architecture—a series of several recurrent layers stacked together—than to simply expand a single layer with greater numbers of cells. The finding supports how deeper model architectures have an advantage when it comes to modeling richer temporal dependencies between features in the UAV audio signals. Both stacked LSTM and

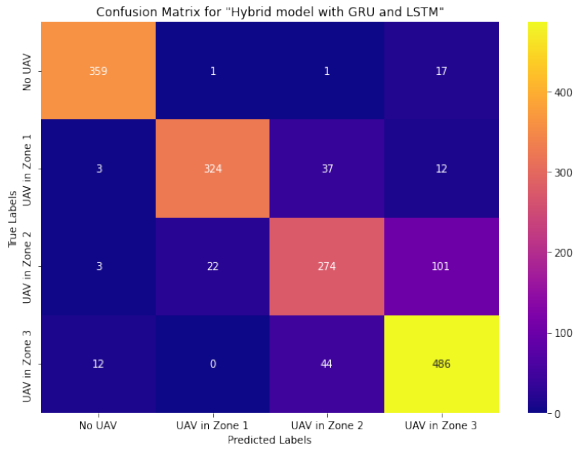
stacked GRU networks successfully learned hierarchical features of the input signals, leading to enhanced classification results. Significantly, recognition accuracy by both LSTM and GRU models with a stacked configuration was almost indistinguishable, which indicates both architectures to be as good at tackling the classification problem with depth augmentation as with expansion of cells. These findings contribute significantly to understanding optimal recurrent network structuring for detection and classification of UAV audio signals.

4.5 Hybrid Deep Neural Network Models

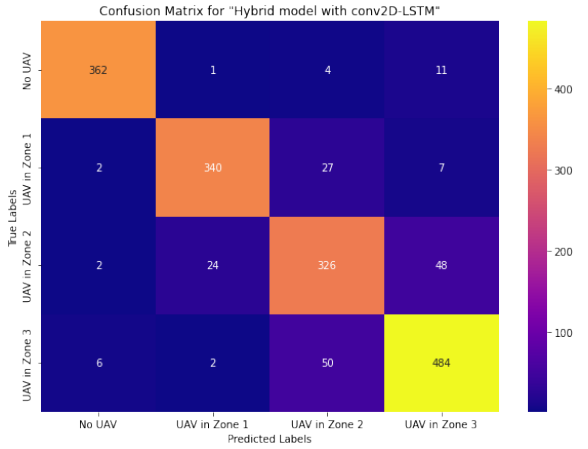
This section discusses the investigation of hybrid neural model structures with an emphasis on improving UAV audio signal recognition. The study starts with an investigation of the hybridization of two recurrent model structures, i.e., LSTM and GRU networks. The two are initially combined to test their combined functionality at detecting both short and long temporal patterns of the data. The combination order is thereafter reversed, with GRU being used first, followed by LSTM, as presented in Figure 4.5, to assess whether order of layering affects recognition performance.
Subsequently, BiLSTM and GRU models are combined to see how bidirectional processing of BiLSTM supplements the simplicity and efficiency of GRU. The process tries to incorporate both backwards and forwards contextual comprehension along with limited computational overhead. In the subsequent stage, attention is shifted to hybridizing recurrent models with CNNs. Both LSTM and BiLSTM networks are separately combined with CNNs to utilize spatial feature extraction properties of CNNs with temporal modeling ability of RNNs.
Finally, a hybrid architecture of higher complexity is explored with a three-way integration of CNN, BiLSTM, and GRU networks. The most comprehensive model tries to integrate spatial and temporal patterns recognition fully by harnessing the unique strengths of all three types of networks. Comparing different hybrid configurations to identify the most resilient and precise solution for classifying UAV sounds under different conditions.
In Figure 4.5, confusion matrices are given as a number of different audio files, providing a clearer picture of how and why the classification occurs. It allows the detailed study of how well a model performs for individual classes by providing both the total number of files classified to a class and misclassified to a class. It provides the number of audio files within confusion matrices to enable an accurate study of where the model performs poorly and where it performs well, and how it can be improved at these specific regions. Table 4.5 shows the same confusion matrices, but this time as percentages. Normalizing to a percentage allows for easier comparison of overall model effectiveness between differing datasets and numbers of classes, and not being biased towards larger datasets or larger numbers of conditions. It provides a clear study of how well a model performs under all conditions, giving a generalized study of results.
Experimental results obtained through hybrid model testing unequivocally confirm the superior performance of the CNN and BiLSTM hybrid in comparison to other hybrid models. The hybrid model is especially impressive as it is able to well utilize both CNN and BiLSTM networks to achieve high recognition accuracy under different conditions. Specifically, for instance, the CNN and BiLSTM hybrid model achieved an impressive 90% recognition accuracy for detecting UAVs at farthest range, an important require-



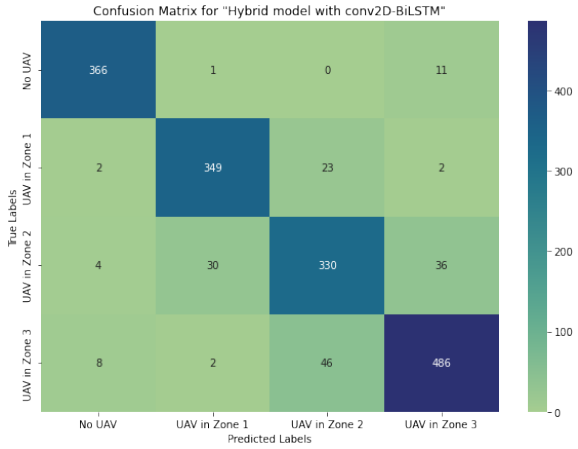
(a) Hybrid model with LSTM and GRU



(b) Hybrid model with GRU and LSTM



(c) Hybrid model with conv2D-LSTM



(d) Hybrid model with conv2D-BiLSTM

Figure 4.5 – Confusion Matrices of Hybrid Deep Learning Architectures

ment for long-range detection purposes. Additionally, for noisy and complex audio data at middle zone, the model also managed to obtain 82% recognition accuracy, thus proving its capability to perform well under suboptimal conditions. The model also performed well at the closest range with an accuracy rate of 93%, successfully dealing with detailed and high-frequency features of close sounds. Finally, one of the most striking results was obtained through this model's capability to recognize and identify background noise with an impressive 97% recognition accuracy. These results validate the power of hybridizing CNN and BiLSTM networks to combat issues of varied distances, varying levels of noise, and complexity of actual audio signals. The capability of the hybrid model to perform with high levels of recognition accuracy for all classes—whether it is recognizing objects at distant distances, differentiating complex acoustic signals under noisy environments, or recognizing background noise—demonstrate the robustness and dependability of this model. In contrast to hybridization with other architectures, CNN-BiLSTM consistently performed at a higher level, especially with regards to stability and accuracy under different conditions. These results demonstrate the practical viability of hybridizing CNN and BiLSTM networks for usage in scenarios where high recognition accuracy is a requirement, for example, in UAV acoustic signal detection, where environmental conditions like noise and varying distances can make it difficult. By successfully managing these issues,

CNN-BiLSTM hybrid model is a good candidate for future research in this field.

Table 4.5 – Evaluation of Model Performance and Predictive 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

It was mainly with an intent to study and compare the recognition abilities of different deep learning model architectures for the specific purpose of prediction of distance of UAVs through sounds. The main intention was to compare and study how well different models did with respect to prediction of accurate distances of the UAVs based on sounds, which is a task that presents multifaceted challenges with differences in sounds, environmental background noise, and differences in distances of the UAVs at detection. With controlled experiments and systematic experiments, the models were tested and thoroughly studied for how well they can handle and interpret distance-correlated sounds and how different environmental factors and noise affect them.
In summary, these experimental results were mainly centered around measuring recognition accuracy and overall prediction performance of all deep learning architectures. These performance measures were crucial to knowing how well a model could recognize and classify UAVs through auditory data and at what distances. By measuring and comparing the performance of different model types, for example, CNNs, RNNs, LSTMs, GRUs, and their hybrids, research attempted to identify the most efficient architectures for this particular task of UAV sound prediction. It not only was focused on measuring raw recognition accuracy but also sought to compare the robustness of models for different situations, for example, with changing background noise and at different distances for the UAV. Additionally, different models and combinations of models that perform best under various circumstances were identified through these experiments, providing an understanding of how various architectures can be utilized for UAV detection in practical

scenarios. The study considered carefully how and why each model's strengths and limitations were influenced by an underlying architecture, including how performance was affected by far-distance detection, noisy environments, and complicated audio signals. Through a detailed comparison of how well each model performed, the research ended up providing guidelines for which deep learning architectures are most appropriate for UAV sound-based distance prediction, an extension to larger disciplines within fields of UAV detection frameworks. Overall, this group of experiments provided an expansive understanding of various deep learning architectures and how and why various architectures have practical utility for UAV recognition and distance estimation through sounds.

Chapter 5

Discussion

In this study, the performance of these deep learning architectures for prediction of UAV sounds at different distances is investigated through extensive empirical evaluations. More specifically, five different methods were tested: convolutional neural networks (CNN), one-layer recurrent neural networks (RNN), two-layer stacked RNN, one-layer RNN with improved cells, voting system, and hybrid models. The goal of this research was to identify how well different types of models predict UAV motions at different range segments, taking into consideration both environmental and engineering conditions of UAV detection.

Within the scope of this research, there were three specific regions where there were to be expected suspicious UAV behaviors, each with its own specific set of difficulties and concerns regarding detection. These regions were determined to be:

Zone 1 – Closest Zone (Up to 15 Meters): It is the zone where drones are at relatively close distances with respect to the detection system. As a result of being close, the received audio signals will generally be clearer, but difficulties will be experienced with sensing rapidly moving drones at these close ranges. Zone 2 – Middle Zone (15 to 30 meters): Within this middle zone, drones remain relatively close but potentially traveling at higher speeds, and there can be a dilution of the audio signal with distance. Models need to be capable of distinguishing between close-range and middle-range drones, as the sound signals will be degrading and potentially corrupted by environmental noise. Zone 3 – Far Zone (30-50 meters): The farthest range of detection of the UAVs under research, where it is harder for sounds to be picked up with accuracy because distance and interferences created by background sounds are more likely to be encountered. The models have to be especially efficient at separating sounds of UAVs from sounds around them, which are likely to be louder at distances. Flight tests were performed at different types of UAV motion patterns, including fully dynamic, partially static, and semi-dynamic patterns of movement. The motion patterns were chosen to simulate actual flight conditions and examine how the models respond under various types of UAV patterns of movement. The speed of the average UAV in these tests was between 3 to 15 meters per second, giving a range of flight conditions between slower-moving and faster, dynamic types of movement. These speeds were utilized to calculate maximum range for all types of zones, with the detection system capable of capturing a maximum range of 50 meters. The UAVs were tested within this range to verify that the detection system was capable of capturing sound signals for all types of zones.

The acoustic data registered during the tests were decomposed into flight parameters of

the drones, enabling a finer understanding of how various flight behaviors contributed to predictability of UAV distance. Such parameters were most important to knowing the difficulties with recognizing UAV sounds because various flight modes produce considerably different sounds that must be properly sensed and interpreted through deep learning frameworks.

Overall, this study comprehensively explored the performance of several deep learning architectures for predicting sounds of UAVs at disparate distances. Through assessing various models and their individual merits and demerits for UAV sound detection, the study contributed to the area of UAV detection systems in several important ways. The main contributions of this study are:

The investigation examined the feasibility of predicting UAV distances through the use of deep learning models that had been trained on sound signals for the purposes of assessing the performance viability of audio-based systems for detecting UAVs for real-world applications. The investigation aimed to identify the accuracy with which different architectures in deep learning could predict UAV proximity under a variety of environmental and operational conditions. By examining a broad sweep of model architectures ranging from Convolutional Neural Networks (CNNs) through Recurrent Neural Networks (RNNs) to Long Short-Term Memory networks (LSTMs), Gated Recurrent Units (GRUs), and hybrid models, the research provided a holistic evaluation of each approach's performance capabilities and limitations.

The testing was conducted over various test zones, divided into close, middle, and far ranges, and included in-situ challenges like environmental noise, random flight patterns, and background interference. These factors were introduced in order to represent the diversity found in real-world environments in which UAV detection systems are generally put into operation, like city environments or surveillance coverage areas.

The results from the experiments clearly affirm that UAV distance prediction with deep learning is not only possible but also exceptionally effective. Of all the models that were experimented with, the hybrid architecture that combined CNN and BiLSTM components was the strongest and most accurate in all zones. This model performed consistently well in recognizing UAV distances, even in noisy and unstructured conditions. The CNN layers performed exceptionally in making inferences from the space and spectrum features from the input sounds, and the BiLSTM layers captured well the temporal dynamics of sound patterns, allowing the hybrid model to make sense of sophisticated sound sequences and recognize UAV proximity with higher accuracy.

Despite the robust performance, the study also revealed a severe limitation: the necessity for large-scale and diverse datasets that can facilitate real-time, deployable UAV detection systems. Although the current realization of the hybrid CNN-BiLSTM model demonstrates considerable potential, real-time application in real-world environments, such as for surveillance, airspace protection, or UAV air traffic control, demands additional data accumulation, performance testing, and system optimization. Without such improvements, the system can fail to generalize well to unseen and novel conditions. Even so, the study establishes a solid basis for future progress in sound-based UAV detection and tracking technologies. It offers insightful conclusions on the merits and demerits of different architectures for deep learning, thus providing a blueprint for improving and fine-tuning the models for practical application. The hybrid CNN-BiLSTM, in specific, offers considerable promise for being put into next-generation UAV surveillance systems. Nonetheless, future efforts in research will be necessary to enhance the robustness, scalability, and versatility of such models in different and changing environments.

This paper meaningfully advances the emerging area of UAV detection employing deep learning, supporting the notion that acoustic data, when combined with the proper model architecture, may be a viable modality for drone surveillance. With refinements and further development, systems like this hold promise as part of the overall surveillance and defense systems.

Chapter 6

Conclusions and future work

6.1 Conclusions

In summary, the current study conducts a comprehensive and in-depth investigation into the performance of different architectures of deep learning on the task of calculating the distances of unmanned aerial vehicles (UAVs), also referred to as drones, based on acoustic signal data. The main emphasis in the study lies with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), two commonly applied classes of neural networks in current deep learning tasks. The investigation was conducted in a manner that sought to compare and contrast the models in the accuracy with which they could estimate UAV distances from a collection of sound recordings taken from a sequence of controlled flight tests. These flight tests were performed in different zones with varying spatial proximities (close, middle, and far), and covered a variety of environmental conditions, including conditions with background noise and with difficult flight maneuvers. These conditions had been specifically included in order to simulate real-world operational environments as well as test the robustness and generalization capability of the models in different and presumably adverse conditions.

Through a rigorous and empirically guided evaluation framework, several leading deep learning architectures, i.e., CNNs, standard RNNs, Long Short-Term Memory networks (LSTMs), and Bidirectional LSTM networks (BiLSTMs), were studied in order to identify which of these configurations is best poised to tackle the intrinsic intricacies of drone-acquired acoustic signatures, which can be quite variable in accordance with distance, type of drone, speed of rotors, and surrounding noise.

The findings from the experimental phase of this study clearly demonstrate that the hybrid architecture with the integration of CNN and BiLSTM layers shows excellent and consistent predictive performance in all the specified distance regions—i.e., the close, middle, and far zones. This hybrid architecture was well suited to tackle the challenges posed by varying distances and dynamic environmental conditions. The feature extractive capability of the CNN part was particularly robust in the form of extractions of both spatial and spectrum features from the sound input, well detecting patterns in the frequency space that form the basis for discerning the characteristic sound signatures of UAVs. Concurrently, the presence of the BiLSTM part served well in detecting temporal dependencies in the sound sequences, allowing the model to understand the evolution of sound patterns with time. This understanding is vital for making sound inferences

in respect to the behavior as well as the movement of the drones, independent of noise pollution or varying closeness.

The integration of both the convolutional and recurrent types of layers enabled a thorough and insightful understanding of the incoming audio data. Not only did this two-level structure increase the accuracy in distance prediction, but also in the system's generalizability to unseen UAVs and new flight environments. It is thus that the hybrid CNN-BiLSTM model comes as a robust and versatile candidate for real-world application, wherein variability and randomness pervades.

In contrast, although the CNN-BiLSTM architecture performed best compared to others in long-range distance estimation tasks, the study also found that models built on recurrent layers alone—such as standard RNNs, stacked RNNs, and GRUs—exhibited worthwhile performance in some specific sub-tasks. These models based on RNNs performed notably well in tasks involving load estimation, as well as on binary classification tasks, especially when the classification task itself was not particularly challenging (e.g., distinguishing if a UAV was in the area or not). Since they perform well in dealing with sequential data, RNNs lend themselves well to time-series processing and performed well in situations where temporal dependencies mattered but lesser spatial complexity was required. In tests that compared the capabilities of single-layer, multi-layer (or stacked), and GRU-augmented RNN models, it was observed that each performed best in various contexts. Despite certain limitations in dealing with high-dimensional and noisy or highly variable acoustic data, the models nonetheless provided adequate performance for binary detection and for inference requirements that need relatively fast and lightweight inference. These results imply that, given the application's specific requirements and constraints—such as real-time responsiveness or computational efficiency—less-complex RNN architectures can remain practically useful, especially in simple classification tasks.

One of the most important takeaways of this research is that there is the potential to predict UAV distance in real-time, even with fairly noisy environments, provided there is adequate and plentiful audio data available for model training. The model accuracy, and specifically that of CNN-BiLSTM, shows that it is indeed possible for audio-based UAV detection systems to perform well under conditions that are similar to real-world conditions. The research, however, presents one of the shortcomings of the approach to date: that the acoustic dataset compiled for experiments was somewhat limited and not much diversely represented with regards to specific numbers of different models of UAV evaluated. The lack of a large variety of types of UAVs being available within the dataset limits how generalizable the results may be to real-world situations where UAVs with various different sounds and flight patterns could be encountered. This constraint is significant to keep in mind when contemplating operational environments' scalability and applicability of the UAV distance prediction system. There exist different shapes and sizes of UAVs, with different acoustic profiles based upon, for example, rotor speed, weight, and flight modes. Therefore, future research will be directed toward increasing the acoustic dataset to cover a larger variety of UAV models, flight modes, and environmental conditions. This will aid in enhancing robustness of deep learning models to be able to process a larger variety of flight conditions and types of drones, which is crucial for real-time deployment across various environments.

In spite of this constraint, outcomes of the study decidedly show the practicability of applying deep learning models to predict UAV distances based on acoustic signals. With advances in deep learning, especially with CNN and BiLSTM model integration, real-time detection and tracking of UAVs are an achievable prospect. With increasing variety

and scope of available acoustic data, these models can be fine-tuned to increase their predictability and accuracy. The study is a critical step toward developing acoustic-detection systems for UAVs, and with more data and fine-tuning, these systems have the potential to be optimally employed across a variety of surveillance, security, and monitoring applications.

Overall, the major contributions of this study are as follows:

The research delves into how deep learning models, especially CNN and RNN-based systems, can be employed to predict UAV distances based on acoustic signals. The CNN-BiLSTM hybrid model was found to provide most accurate and consistent predictions, especially while overcoming difficulties that arise with varying UAV distances and environmental noise. RNN models exhibited good performance in easier classification and estimation of loads, giving good indication of how sequential models can be applied to detect UAVs.

The research points to the applicability of real-time prediction of distances via UAV, with a recognition for the necessity of a larger acoustic dataset to make the system applicable and scalable to a wider scope. Future steps will include collecting a larger and more diversified set of audio data, including an expanded set of UAV types, flight regimes, and noise levels, to increase the robustness of prediction models. Future research could also investigate how to integrate additional sensor modalities, including visual data streams from cameras or radar signals, into detection systems to further increase the accuracy and dependability of UAV detection. Integrating various types of sensors with deep learning models presents a promising future area of research, potentially enabling multi-modal detection systems with enhanced capability to detect and track UAVs in ever-more sophisticated and dynamic environments. Eventually, as this research has proven promising for deep learning for prediction of distances for UAVs, there is a need for future research to improve these models to make them applicable for real-time usage in various practical environments. Nevertheless, the results build a firm basis for future developments of autonomous systems for UAV detection, and offer insightful understanding of challenges and possibilities for them.

6.2 Limitations

This study has a number of drawbacks even if it provides insightful information about UAV detection and distance estimation by deep learning-based audio analysis. The models' capacity to be applied to a variety of real-world situations may be constrained by the data's collection under a narrow range of environmental conditions. The difficulty of long-range detection is further highlighted by the fact that sound attenuation and background noise cause the system's performance to decline with increasing distance. The quality of the characteristics that are collected may be impacted by the fidelity limits that come with using typical laptop microphones. Furthermore, whereas hybrid and ensemble models improved accuracy, they also made real-time implementation on devices with limited resources less possible due to their increasing computational complexity. Bias and a lack of granularity in model learning may have been induced by class imbalance and wide UAV category labelling. Lastly, the system solely relies on audio input, which could not be enough in acoustic contexts that are noisy or unclear, indicating the need for multimodal techniques in the future.

6.3 Future work

Future studies will focus on substantially augmenting the acoustic dataset to increase the robustness, flexibility, and overall accuracy of UAV identification systems. By being fed a greater and more heterogeneous pool of sound samples, the models will be familiarized with a wider diversity of UAV sounds, including variations caused by various drone models, motor configurations, flight altitudes, speeds, environmental conditions, and flight patterns. This more comprehensive dataset will allow the deep learning models to generalize better, minimizing overfitting to the limited situations present in the current data. This extended data addresses already revealed limitations—most notably, the limited number of UAV types included in the original dataset—which has a direct bearing on the model's performance in real-world scenarios. Drones of different sizes, configurations, and sonic signatures will be taken into account, including hobby drones, commercial UAVs, and stealthy surveillance drones, each with distinct sonic signatures that must be learned by the system.

With a wider, more diverse dataset, the acoustic model will be in a stronger position to identify UAVs in a broad spectrum of operational environments, facilitating better distance estimation in real-world conditions. Environmental factors like wind, urban noise, and terrain, for instance, can dramatically impact acoustic perception. By training on samples obtained in a variety of environments, both urban and rural, the model's capacity to function reliably in noisy, cluttered, or visibility-constrained environments will be markedly improved. This is necessary for guaranteeing robust UAV detection performance in real-world scenarios like public event monitoring, border patrol, or sensitive facility guarding.

In addition to dataset expansion, future studies also pursue the integration with other sensor modalities into the audio-based recognition system in order to develop a multimodal UAV detection framework. This integration has the potential to enhance system performance by taking advantage of complementary information. Observations based on visual data in the form of RGB or thermal cameras can glean valuable information on the physical position, path, and orientation of the UAV, as radar systems offer accurate measurements for speed, direction, and even altitude for the UAV. This sensor data fusion, in which multiple sensor data is combined—also known as sensor fusion—has the potential to overcome the weaknesses of single modalities. For example, as winds or city noise could momentarily degrade the quality of the necessary audio signals, visual or radar sensors can remain unaffected. On the flip side, in cases like fog, smoke, or nighttime, where visual systems are disabled, acoustic data can be unblemished.

The emergence of such bimodal or multimodal systems is a step forward toward real-time, intelligent drone surveillance systems. They can provide improved levels of situation awareness and decision-making, necessary for numerous critical applications ranging from military surveillance to law enforcement, border patrol, disaster relief, and environmental surveillance. The application of deep learning algorithms for processing and interpreting multimodality in real time is going to be instrumental in improving detection accuracy, eliminating false alarms, and facilitating predictive analytics. For instance, a multimodal system could be able to anticipate a drone's future path based on previous patterns of sound, visual tracking data, and radar velocity data, allowing for reactive countermeasures.

Moreover, the systems can be created in a scalable form that can be deployed in different environments such as stationary monitor stations, mobile ground, or even air-based

platforms. This scalability gives the opportunity for large-scale rollout in a variety of industries ranging from agricultural or infrastructure inspections to search-and-rescue scenarios as well as wildlife tracking. The system architecture can also be made flexible through modularity so that it can be easily upgraded with emerging sensor technologies as they become viable, thereby maintaining long-term sustainability and ongoing performance enhancement.

The study also highlights the need for training these models with high-quality, diverse data, which is directly obtained in representative real-world environments. Not only will this increase the model's accuracy and robustness, but it also offers valuable information on how such systems can be deployed in moderately noisy environments. High-noise environments may be challenging for systems based on sound, but they can be effectively dealt with in the presence of proper data pre-processing techniques, denoising algorithms, and multimodal sensor approaches.

Eventually, the infusion of smart prediction models, backed by sophisticated machine learning and deep learning methods, is expected to dramatically improve the functionality and reliability of UAV detection systems. In the future, such systems are bound to be highly autonomous, with the capacity for ongoing learning, flexible behavior, and real-time decision-making with less or no supervision from humans. These technologies will be critical for addressing increasing worldwide demand for safe, efficient drone surveillance systems, guaranteeing safety, compliance, and efficiency in a broad spectrum of sectors.

Overall, the future vision outlined calls for a comprehensive advancement of drone detection technologies, from starting with dataset expansion through to the realization of intelligent, scalable, multimodal systems that can operate in real-world environments. The method promises to bring scientific concepts in acoustic drone detection from the laboratory into practical reality with far-reaching social and industrial implications.

Bibliography

[1] S. Li, H. Kim, S. Lee, J. C. Gallagher, D. Kim, S. Park, and E. T. Matson, "Convolutional neural networks for analyzing unmanned aerial vehicles sound," in *2018 18th International Conference on Control, Automation and Systems (ICCAS)*, 2018, pp. 862–866.

[2] E. Bowman, "Canada beats new zealand in women's soccer as olympic spy drone scandal grows," in *WVIA Radio*, July 25, 2024. [Online]. Available: https://www.wvia.org/news/2024-07-25/canada-beats-new-zealand-in-womens-soccer-as-olympic-spy-drone-scandal-grows

[3] H. Kesteloo, "Drone drama at olympics: Canada accused of spying on new zealand soccer team," in *DroneXL*, July 24, 2024. [Online]. Available: https://www.rnz.co.nz/news/olympics-2024/523172/olympics-canada-beat-football-ferns-amid-spying-scandal

[4] "Man arrested for flying drone over the spacex facility," in *C-UAS Hub*, 2024. [Online]. Available: https://cuashub.com/en/content/man-arrested-for-flying-drone-over-the-spacex-facility/

[5] D. Utebayeva, L. Ilipbayeva, and E. T. Matson, "Practical study of recurrent neural networks for efficient real-time drone sound detection: A review," *Drones*, vol. 7, no. 1, 2023. [Online]. Available: https://www.mdpi.com/2504-446X/7/1/26

[6] U. Seidaliyeva, D. Akhmetov, L. Ilipbayeva, and E. T. Matson, "Real-time and accurate drone detection in a video with a static background," *Sensors*, vol. 20, no. 14, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/14/3856

[7] U. Seidaliyeva, L. Ilipbayeva, K. Taissariyeva, N. Smailov, and E. T. Matson, "Advances and challenges in drone detection and classification techniques: A state-of-the-art review," *Sensors*, vol. 24, no. 1, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/1/125

[8] U. G. A. Office, "Counter-uas: Actions needed to improve dod's efforts to address unmanned aircraft systems," https://www.gao.gov/products/gao-22-105844, 2022, accessed: 2025-06-11.

[9] D.-F. Solutions, "Counter-drone incidents: Real-world events involving rogue uavs," https://www.d-fendsolutions.com/counter-drone-incident-reports/, 2025, accessed: 2025-06-11.

[10] A. Press, "Drone disrupts public event, raising safety concerns," https://apnews.com/article/drone-public-event-safety, 2025, accessed: 2025-06-11.

[11] S. Ding, X. Guo, T. Peng, X. Huang, and X. Hong, "Drone detection and tracking system based on fused acoustical and optical approaches," *Advanced Intelligent Systems*, 2023, first published: 19 July 2023. [Online]. Available: https://doi.org/10.1002/aisy.202300251

[12] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–8.

[13] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, and H. G. Okuno, "Design of uav-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, p. 2535, 2017.

[14] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic source localization from multirotor uavs," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, pp. 8618–8628, 2019.

[15] S. Wu, Y. Zheng, K. Ye, H. Cao, X. Zhang, and H. Sun, "Sound source localization for unmanned aerial vehicles in low signal-to-noise ratio environments," *Remote Sensing*, vol. 16, no. 11, p. 1847, 2024.

[16] S. V. Sibanyoni, D. T. Ramotsoela, B. J. Silva, and G. P. Hancke, "A 2-d acoustic source localization system for drones in search and rescue missions," *IEEE Sensors Journal*, vol. 19, no. 1, pp. 332–341, 2018.

[17] G. Jekateryńczuk and Z. Piotrowski, "A survey of sound source localization and detection methods and their applications," *Sensors*, vol. 24, no. 1, p. 68, 2023.

[18] L. Kraljević, M. Russo, M. Stella, and M. Sikora, "Free-field tdoa-aoa sound source localization using three soundfield microphones," *IEEE access*, vol. 8, pp. 87 749–87 761, 2020.

[19] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[20] D. Utebayeva and A. Yembergenova, "Study a deep learning-based audio classification for detecting the distance of uav," in *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2024, pp. 1–7.

[21] D. Lim, H. Kim, S. Hong, S. Lee, G. Kim, A. Snail, L. Gotwals, and J. C. Gallagher, "Practically classifying unmanned aerial vehicles sound using convolutional neural networks," in *2018 Second IEEE International Conference on Robotic Computing (IRC)*, 2018, pp. 242–245.

[22] W. Cai, L. Wu, Y. Cui, and S. He, "Uncertainty principle and power quality sensing and analysis in smart substation," *Sensors*, vol. 20, no. 15, p. 4281, 2020. [Online]. Available: https://doi.org/10.3390/s20154281

[23] A. Calanca and P. Fiorini, "Impedance control of series elastic actuators based on well-defined force dynamics," *Robotics and Autonomous Systems*, vol. 96, pp. 69–81, 2017. [Online]. Available: https://doi.org/10.1016/j.robot.2017.06.013

[24] M. Nijim and N. Mantrawadi, "Drone classification and identification system by phenome analysis using data mining techniques," in *2016 IEEE Symposium on Technologies for Homeland Security (HST)*, 2016, pp. 1–5.

[25] DJI, "Phantom 4 - product information and support," 2020, accessed: 2025-06-11. [Online]. Available: https://www.dji.com/support/product/phantom-4

[26] Intel Corporation, "Intel falcon 8+ drone specifications," 2020, accessed: 2025-06-11. [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/98476/intel-falcon-8/specifications.html

[27] DJI, "Spreading wings s1000 - product page," 2020, accessed: 2025-06-11. [Online]. Available: https://www.dji.com/spreading-wings-s1000

[28] A. Sedunov, D. Haddad, H. Salloum, A. Sutin, N. Sedunov, and A. Yakubovskiy, "Stevens drone detection acoustic system and experiments in acoustics uav tracking," in *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, 2019, pp. 1–7.

[29] A. Cheranyov and E. Dukhan, "Methods of detecting small unmanned aerial vehicles," in *2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, 2021, pp. 0218–0221.

[30] U. Papa, G. Del Core, G. Giordano, and S. Ponte, "Obstacle detection and ranging sensor integration for a small unmanned aircraft system," in *2017 IEEE International Workshop on Metrology for AeroSpace (MetroAeroSpace)*, 2017, pp. 571–577.

[31] B. Jang, Y. Seo, B. On, and S. Im, "Euclidean distance based algorithm for uav acoustic detection," in *2018 International Conference on Electronics, Information, and Communication (ICEIC)*, 2018, pp. 1–2.

[32] L. Jiqing, F. Husheng, Y. Qin, and Z. Chunhua, "Quad-rotor uav audio recognition based on mel spectrum with binaural representation and cnn," in *2021 International Conference on Computer Engineering and Application (ICCEA)*, 2021, pp. 285–290.

[33] J. Kim, C. Park, J. Ahn, Y. Ko, J. Park, and J. C. Gallagher, "Real-time uav sound detection and analysis system," in *2017 IEEE Sensors Applications Symposium (SAS)*, 2017, pp. 1–5.

[34] B. Yang, E. T. Matson, A. H. Smith, J. E. Dietz, and J. C. Gallagher, "Uav detection system with multiple acoustic nodes using machine learning models," in *2019 Third IEEE international conference on robotic computing (IRC)*. IEEE, 2019, pp. 493–498.

[35] M. Z. Anwar, Z. Kaleem, and A. Jamalipour, "Machine learning inspired sound-based amateur drone detection for public safety applications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2526–2534, 2019.

[36] F. R. E. Fagiani, "Uav detection and localization system using an interconnected array of acoustic sensors and machine learning algorithms," Master's thesis, Purdue University, 2021.

[37] C. A. Ahmed, F. Batool, W. Haider, M. Asad, and S. H. R. Hamdani, "Acoustic based drone detection via machine learning," in *2022 International Conference on IT and Industrial Technologies (ICIT)*. IEEE, 2022, pp. 01–06.

[38] D. Tejera-Berengue, F. Zhu-Zhou, M. Utrilla-Manso, R. Gil-Pita, and M. Rosa-Zurera, "Acoustic-based detection of uavs using machine learning: Analysis of distance and environmental effects," in *2023 IEEE Sensors Applications Symposium (SAS)*. IEEE, 2023, pp. 1–6.

[39] S. Salman, J. Mir, M. T. Farooq, A. N. Malik, and R. Haleemdeen, "Machine learning inspired efficient audio drone detection using acoustic features," in *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*. IEEE, 2021, pp. 335–339.

[40] M. Ohlenbusch, A. Ahrens, C. Rollwage, and J. Bitzer, "Robust drone detection for acoustic monitoring applications," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 6–10.

[41] E. R. Solis, D. V. Shashev, and S. V. Shidlovskiy, "Implementation of audio recognition system for unmanned aerial vehicles," in *2021 International Siberian Conference on Control and Communications (SIBCON)*. IEEE, 2021, pp. 1–8.

[42] H. Lee, S. Han, J.-I. Byeon, S. Han, R. Myung, J. Joung, and J. Choi, "Cnn-based uav detection and classification using sensor fusion," *IEEE Access*, vol. 11, pp. 68 791–68 808, 2023.

[43] I. Ku, S. Roh, G. Kim, C. Taylor, Y. Wang, and E. T. Matson, "Uav payload detection using deep learning and data augmentation," in *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*, 2022, pp. 18–25.

[44] N. Nakngoen, P. Pongboriboon, N. Inthanop, J. Akharachaisirilap, T. Woodward, and N. Teerasuttakorn, "Drone classification using gated recurrent unit," in *IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society*, 2023, pp. 1–4.

[45] E. R. Solis, D. V. Shashev, and S. V. Shidlovskiy, "Implementation of audio recognition system for unmanned aerial vehicles," in *2021 International Siberian Conference on Control and Communications (SIBCON)*, 2021, pp. 1–8.

[46] P. Racinskis, J. Arents, and M. Greitans, "(poster) drone detection and localization using low-cost microphone arrays and convolutional neural networks," in *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, 2023, pp. 80–82.

[47] B. Kim, B. Jang, D. Lee, and S. Im, "Cnn-based uav detection with short time fourier transformed acoustic features," in *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, 2020, pp. 1–3.

[48] S. Jeon, J.-W. Shin, Y.-J. Lee, W.-H. Kim, Y. Kwon, and H.-Y. Yang, "Empirical study of drone sound detection in real-life environment with deep neural networks," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1858–1862.

[49] S. Al-Emadi, A. Al-Ali, and A. Al-Ali, "Audio-based drone detection and identification using deep learning techniques with dataset enhancement through generative adversarial networks," *Sensors*, vol. 21, no. 15, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/15/4953

[50] P. Casabianca and Y. Zhang, "Acoustic-based uav detection using late fusion of deep neural networks," *Drones*, vol. 5, no. 3, 2021. [Online]. Available: https://www.mdpi.com/2504-446X/5/3/54

[51] İlhan Aydın and E. Kızılay, "Development of a new light-weight convolutional neural network for acoustic-based amateur drone detection," *Applied Acoustics*, vol. 193, p. 108773, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X22001475

[52] C. Dumitrescu, M. Minea, I. M. Costea, I. Cosmin Chiva, and A. Semenescu, "Development of an acoustic system for uav detection," *Sensors*, vol. 20, no. 17, p. 4870, 2020.

[53] H. C. Vemula, "Multiple drone detection and acoustic scene classification with deep learning," Master's thesis, Wright State University, 2018.

[54] Y. Wang, Z. Chu, I. Ku, E. C. Smith, and E. T. Matson, "A large-scale uav audio dataset and audio-based uav classification using cnn," in *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2022, pp. 186–189.

[55] S. S. Katta, S. Nandyala, E. K. Viegas, and A. AlMahmoud, "Benchmarking audio-based deep learning models for detection and identification of unmanned aerial vehicles," in *2022 Workshop on Benchmarking Cyber-Physical Systems and Internet of Things (CPS-IoTBench)*. IEEE, 2022, pp. 7–11.

[56] B. Taha and A. Shoufan, "Machine learning-based drone detection and classification: State-of-the-art in research," *IEEE access*, vol. 7, pp. 138 669–138 682, 2019.

[57] X. Mu, J. Lu, P. Watta, and M. H. Hassoun, "Weighted voting-based ensemble classifiers with application to human face recognition and voice recognition," in *2009 International Joint Conference on Neural Networks*, 2009, pp. 2168–2171.

[58] M. H. Moattar and M. M. Homayounpour, "A weighted feature voting approach for robust and real-time voice activity detection," *ETRI Journal*, vol. 33, no. 1, pp. 99–109, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.4218/etrij.11.1510.0158

[59] H. Zhang and Y. Zhou, "A neural network-based weighted voting algorithm for multi-target classification in wsn," *Sensors*, vol. 24, no. 1, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/1/123

[60] S. Liang and H. Chongzhao, "Dynamic weighted voting for multiple classifier fusion: A generalized rough set method," *Journal of Systems Engineering and Electronics*, vol. 17, no. 3, pp. 487–494, 2006.

[61] X. Wang, B. Ma, Z. Yu, F. Li, and Y. Cai, "Multi-scale decision network with feature fusion and weighting for few-shot learning," *IEEE Access*, vol. 8, pp. 92 172–92 181, 2020.

[62] O. Gokalp and E. Tasci, "Weighted voting based ensemble classification with hyper-parameter optimization," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2019, pp. 1–4.

[63] L. Nanni, M. Paci, G. Maguolo, S. Ghidoni, and S. Brahnam, "Audio classification," *Applied Sciences*, vol. 11, no. 13, p. 5796, 2021. [Online]. Available: https://doi.org/10.3390/app11135796

[64] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[65] K. Choi, G. Fazekas, and M. Sandler, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[66] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6645–6649.

[67] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014. [Online]. Available: https://arxiv.org/abs/1412.3555

[68] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and K. Yoshii, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[69] M. Müller, "Fundamentals of music processing: audio, analysis, algorithms, applications," 2015.

[70] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[71] G. Kim and H. K. Kim, "Drone detection using sound based deep neural network with complementary features," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 2017, pp. 837–841. [Online]. Available: https://dl.acm.org/doi/10.1145/3052973.3053002

[72] DJI, "Dji - official website," 2025, accessed: 2025-06-11. [Online]. Available: https://www.dji.com/global

[73] U. Seidaliyeva, L. Ilipbayeva, K. Taissariyeva, N. Smailov, and E. T. Matson, "Advances and challenges in drone detection and classification techniques: A state-of-the-art review," *Sensors*, vol. 24, no. 1, p. 125, 2024. [Online]. Available: https://doi.org/10.3390/s24010125

[74] D. Utebayeva, M. Alduraibi, L. Ilipbayeva, and Y. Temirgaliyev, "Stacked bilstm - cnn for multiple label uav sound classification," in *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, 2020, pp. 470–474.

[75] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Proc. Interspeech*, 2015, pp. 1–5, https://www.isca-speech.org/archive/interspeech$_2$015/$sainath15_interspeech.html$.

[76] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream cnn based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, 2019. [Online]. Available: https://doi.org/10.3390/s19071733

[77] D. Utebayeva, L. Ilipbayeva, U. Seidaliyeva, A. Yembergenova, and E. T. Matson, "Deep learning models for predicting drone sound distances: Lightweight, fusion and hybridization approaches," *Preprints*, 2024. [Online]. Available: https://doi.org/10.20944/preprints202410.2156.v1

[78] D. Utebayeva, L. Ilipbayeva, and E. T. Matson, "Practical study of recurrent neural networks for efficient real-time drone sound detection: A review," *Drones*, vol. 7, no. 1, p. 26, 2023. [Online]. Available: https://doi.org/10.3390/drones7010026

[79] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[80] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[81] J. Li, X. Zhang, M. Sun, X. Zou, and C. Zheng, "Attention-based lstm algorithm for audio replay detection in noisy environments," *Applied Sciences*, vol. 9, no. 8, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/8/1539

[82] P. Mobtahej, X. Zhang, M. Hamidi, and J. Zhang, "Combining audio and visual speech recognition using lstm and deep convolutional neural network," *International Journal of Information Technology*, vol. 14, 2022.

[83] I. Lezhenin, N. Bogach, and E. Pyshkin, "Urban sound classification using long short-term memory neural network," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2019, pp. 57–60.

[84] J. Liu, Y. Yin, H. Jiang, H. Kan, Z. Zhang, P. Chen, B. Zhu, and Z. Wang, "Bowel sound detection based on mfcc feature and lstm neural network," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2018, pp. 1–4.

[85] J. M. Navarro, R. Martínez-España, A. Bueno-Crespo, R. Martínez, and J. M. Cecilia, "Sound levels forecasting in an acoustic sensor network using a deep neural network," *Sensors*, vol. 20, no. 3, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/3/903

[86] J. Zhang, Y. Chen, N. Li, J. Zhai, Q. Han, and Z. Hou, "A weak fault identification method of micro-turbine blade based on sound pressure signal with lstm networks," *Aerospace Science and Technology*, vol. 136, p. 108226, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1270963823001232

[87] Y. Huo, M. Jin, and S. You, "Lstm-based framework for the synthesis of original soundtracks," *IEEE Access*, vol. 12, pp. 33 832–33 842, 2024.

[88] R. Lu and Z. Duan, "Bidirectional gru for sound event detection," in *Detection and Classification of Acoustic Scenes and Events 2017*, 2017. [Online]. Available: https://dcase.community/documents/challenge2017/technical_reports/DCASE2017_Lu_137.pdf

[89] J. Saputra, A. Prasetiadi, and A. Iqsyahiro Kresna, "Generating background music from vocal sound with low-end hardware using autoencoder and gru," in *2023 International Electronics Symposium (IES)*, 2023, pp. 428–433.

[90] M. Soundarya and S. Anusuya, "An investigational analysis of automatic speech recognition on deep neural networks and gated recurrent unit model," in *Advances in Data-Driven Computing and Intelligent Systems*, S. Das, S. Saha, C. A. Coello Coello, and J. C. Bansal, Eds. Singapore: Springer Nature Singapore, 2024, pp. 45–60.

[91] N. Kulshrestha, "Use of deep learning methods such as lstm and gru in polyphonic music generation," Ph.D. dissertation, Dublin, National College of Ireland, 2020.

[92] J. Rusrus, S. Shirmohammadi, and M. Bouchard, "Characterization of moving sound sources direction-of-arrival estimation using different deep learning architectures," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–14, 2023.

[93] S. Chandratre, P. Warule, S. P. Mishra, and S. Deb, "Lstm-and gru-based common cold detection from speech signal," in *Conference on Frontiers of Research in Speech and Music*. Springer, 2011, pp. 439–448.

[94] H.-G. Kim and J. Y. Kim, "Acoustic event detection in multichannel audio using gated recurrent neural networks with high-resolution spectral features," *ETRI Journal*, vol. 39, no. 6, pp. 832–840, 2017.

[95] E. Fanioudakis and A. Vafeiadis, "Investigating temporal and spectral sequences combining gru-rnns for acoustic scene classification," *the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep*, 2020.

[96] J.-Y. Liu and Y.-H. Yang, "Dilated convolution with dilated gru for music source separation," *arXiv preprint arXiv:1906.01203*, 2019.

[97] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," *arXiv preprint arXiv:1706.09559*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1706.09559

[98] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1506.00019

[99] Wikipedia contributors, "Recurrent neural network," https://en.wikipedia.org/wiki/Recurrent_neural_network, 2024, accessed: June 2025.

[100] P. Casabianca and Y. Zhang, "Acoustic-based uav detection using late fusion of deep neural networks," *Drones*, vol. 5, no. 3, p. 54, 2021. [Online]. Available: https://doi.org/10.3390/drones5030054

[101] Q. Dong, Y. Liu, X. Liu, H. Cao, C. Shen, and J. Tang, "Drone sound detection system based on feature result-level fusion using deep learning," *Multimedia Tools and Applications*, vol. 82, pp. 149–171, 2023. [Online]. Available: https://doi.org/10.1007/s11042-022-12964-3

[102] B. Taha and A. Shoufan, "Machine learning-based drone detection and classification: State-of-the-art in research," *IEEE Access*, vol. 7, pp. 138 669–138 682, 2019.

[103] D. Floreano and R. J. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, no. 7553, pp. 460–466, 2015. [Online]. Available: https://www.nature.com/articles/nature14542