UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science

Roman Karpenko

Drone sound recognition system using machine learning

Drooni heli tuvastamine masinõpe abil riigikaitseks

Master's Thesis (30 ECTS)

Supervisor: Anna Aljanaki

Drone sound recognition system using machine learning

Abstract:

Radar-based detection systems are expensive, and it is currently not possible to cover all of the combat territory with them. However, modern warfare employs numerous aerial strikes with drones. In this thesis, we develop a proof-of-concept system that can detect Shaheed drones and distinguish their sound from airplane and helicopter sounds. Shaheed drones are used in the Russo-Ukrainian war against Ukraine. We test several approaches and find that specific drone can be identified well enough using only sound.

Keywords: machine learning, military intelligence, sound processing, machine listening

CERCS: P176 Artificial intelligence

Drooni heli tuvastamine masinõpe abil riigikaitseks

Lühikokkuvõte: Radaripõhised lennuki tuvastussüsteemid on kallid ja seega ei ole võimalust katta nendega terve lahinguala. Kaasaegses sõjas kasutatakse aga arvukalt õhulööke odavate droonidega. Selles diplomitöös kirjeldame süsteemi prototüüpi, mis suudab tuvastada Shaheedi droone, mida kasutatakse Venemaa-Ukraina sõjas Ukraina vastu. Testime mitmeid lähenemisviise. Testid näitasid et kuigi drooni heli saab piisavalt edukalt eristada lennuki ja helikopteri helist.

Võtmesõnad: masinõpe, sõjaväeluure, helitöötlus, masin kuulamine

CERCS: P176 Tehisintellekt

Table of Contents

Introduction

The goal of this thesis is to develop an early alert system capable of timely notifying soldiers about approaching enemy drones in combat zones. The Russo-Ukrainian War of 2022 introduced an urgency in developing such solutions and was the main reason why we started developing it. Traditional radar systems, often hindered by terrain and constrained by the size of operational areas, face challenges in detecting small, agile targets like drones. Our solution utilizes sound-based drone detection, offering a more responsive and accurate alternative to human detection and, in theory, a much cheaper solution than a radar. Conventional radars are very efficient and robust solutions for tracking the aerial intrusion overall, but due to their cost, mass production of them could not be launched to cover huge territories. In contrast, a **portable smart system consisting only of several microphones** located in a particular distance from each other could be a nice and quick solution for this detection problem.

In September 2022, Russian military started to use Shaheed-136 drones to attack Ukrainian infrastructure and cities. Since Ukrainian forces don't use Shaheed-136 drones, the sound they make is specific to enemy forces and can be used to identify enemy attacks.

Such a system could consist of two modules: one for sound anomaly detection, and another for anomaly classification. A system monitors its environment for acoustical anomalies, records sound, evaluates whether unusual sound is Shaheed-135, and outputs a probability of attack. In addition, a system needs to report where the attack is coming from, which means we need a localization system as well.

The solution developed in this thesis was created as a teamwork between Ukrainian military representatives, who helped with sound recording on a battlefield and with developing and testing a hardware prototype, and the author, who was responsible for the ML prototype.

The thesis is structured to provide a comprehensive overview of the development and implementation of the sound detection system. It begins with a background on signal processing and audio analysis. We also briefly review how machine learning is used for sound recognition. In the next chapter we describe the data collection process, including filtering external audio datasets for recordings and scraping YouTube for relevant data. In the next chapter, the implementation of the system is discussed, describing LSTM, and CNN based approaches. Also, the hardware parts that were used in creation of the test system are briefly discussed. Subsequent sections present the conclusions and discussion.

1 Terms and Notations

Audio Classification	The process of identifying and categorizing segments of audio into different classes based on their characteristics.
Convolutional Neural Network (CNN)	A type of artificial neural network commonly applied to image recognition, image classification, and object detection. CNNs are designed to automatically and adaptively learn spatial hierarchies of features from input data, which allows CNNs to effectively capture patterns and structures within images.
Cross Entropy Loss	A loss function used in machine learning that quantifies the difference between two probability distributions - the actual output and the predicted output by the model.
Data Augmentation (in ML context)	A technique that helps enhance the diversity of the dataset and improve model generalization, which involves creating new training data by applying transformations. For instance, rotation, flipping, scaling, or adding noise, in case of images.
Decision Tree	A machine learning algorithm that partitions the input space into subsets based on the values of input features, with each node representing a decision based on a feature value. Decision trees are used for classification and regression tasks and are interpretable.
Deep learning	Deep learning is a subset of machine learning that utilizes artificial neural networks with many layers (hence "deep") to learn from large amounts of data. Each layer extracts increasingly abstract features from the input data, allowing the network to learn complex representations. Methods used can be supervised, semi-supervised or unsupervised.[34]
Fine Tuning	An approach to transfer learning in which the parameters of a pre-trained model are tuned on new data with a small learning rate. [11] Fine-tuning can be done on the entire neural network, or on only a subset of its layers, in which case the layers that are not being fine-tuned are "frozen" (not updated during the backpropagation step). [46]
Long Short-Term Memory (LSTM)	LSTM stands for Long Short-Term Memory, which is a type of recurrent neural network (RNN) architecture designed to capture long-term dependencies in sequential data.
Machine learning (ML)	A field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions.[32]
Miltech	Military Technology, refers to the advanced technological tools and systems used in military applications.
Overfitting	A modeling error in machine learning where a function is too closely aligned to a

	limited set of data points and fails to generalize well to new data.
Recurrent Neural Network (RNN)	RNN, or Recurrent Neural Network, is a type of artificial neural network designed to process sequential data by maintaining a state or memory of previous inputs. Unlike feedforward neural networks, which process data in a fixed sequence, RNNs have connections that form a directed cycle, allowing them to exhibit dynamic temporal behavior.
Regularization	A technique in machine learning that constrains or regularizes the model's learning capacity to prevent overfitting.
ROC curve (receiver operating characteristic curve)	A graph showing the performance of a classification model at all classification thresholds [77]
Spectrogram	A spectrogram is a visual representation of the spectrum of frequencies in a signal as it varies with time. It is a three-dimensional plot where the x-axis represents time, the y-axis represents frequency, and the color intensity (or brightness) represents the strength or amplitude of the signal at each time-frequency point. [69]
Time difference of arrival (TDOA)	A difference between the absolute time instants, between time of transmission (when a radio signal (electromagnetic impulse) emanates from a transmitter) and time of arrival (when it reaches the receiver)
Transfer learning	The ability of a system to recognize and apply knowledge and skills gained in previous tasks to new tasks [44]
VGGish	A deep convolutional neural network developed by Google, specifically designed for audio feature extraction. It is based on the VGG architecture, which is well-known for its effectiveness in image recognition tasks.
XGBoost	An optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework.
YouTube Scraping	The process of extracting data or content from YouTube videos. In this context, it refers to collecting specific audio data from YouTube content.

2 Background

In an active war zone, a lot is going on in the air. Planes, missiles, drones are flying both from enemy positions and towards them. A lot of noise is produced by explosions, vehicle engines on the ground and in the air. In this chapter, we will explain which are the main aerial targets in Russo-Ukrainian war that are likely to be detected in the air, and how they are usually detected.

The main way to detect objects in the air in modern warfare is radar. Radar detection uses **electromagnetic waves** to identify and track objects. An **active radar** emits radio waves that reflect off objects and return to the receiver, providing information about the object's distance, location, and speed [68]. **Passive radar**, on the other hand, relies on ambient radio signals to detect reflections from objects, making it distinct from active radar systems that emit their own signals [68]. Active radar offers more precise information and is particularly useful for long-range object detection. Active radar is widely used for aircraft detection due to its ability to provide precise information about an aircraft's location and movement. By emitting radio waves that reflect off the aircraft and return to the radar receiver, it provides data on the distance, speed, and trajectory of the aircraft [68]. This makes this type of radar detection a powerful tool for tracking aerial vehicles over long and middle ranges.

Sound waves can also be used for detection purposes, particularly for aerial objects like drones. Sound waves are longitudinal waves that travel by compressing and decompressing a medium. They require a medium for propagation, such as air, water, or a solid, making them distinct from electromagnetic waves that can travel through a vacuum. The speed of sound depends on the medium through which it travels, with sound traveling at approximately 343 meters per second in air at 20°C. This speed can vary with temperature, pressure, or the specific medium being traversed [78].

The range at which a drone can be detected by sound depends on factors such as its volume and frequency output, environmental noise levels, and the sensitivity of the detection equipment [68]. Sound detection can also offer unique insights, including information on a drone's operational status or specific model based on its acoustic signature [41].

Radar provides **more reliable** long-range detection capabilities compared to sound detection, which is limited by attenuation over distance. However, sound detection can operate **passively**, not emitting signals that might alert targets or interfere with other systems. Which makes sound detection a useful complement to radar detection, providing additional information on a drone's presence and status [63].

After we have understood the crucial differences between the sound and radar operation, now we need to understand the definition and classification of the drones.

The drone is an uncrewed aerial vehicle, also known as unmanned aerial vehicle (UAV) [76]. They were first developed in the 1990s to be used in operations where the crew could be severely affected and the operations were too dangerous for them. There are various classifications of drones used in the military, one of them is US Department of Defense (DoD) classification [75], given in Table 1.

Category	Size	Maximum Gross Takeoff Weight (MGTW) (lbs.)	Normal Operating Altitude (ft)	Airspeed (knots)
Group 1	Small	0-20	<1,200 AGL*	<100
Group 2	Medium	21-55	<3,500	<250
Group 3	Large	<1320	<18,000 MSL**	<250
Group 4	Larger	>1320	<18,000 MSL	Any airspeed
Group 5	Largest	>1320	>18,000	Any airspeed
*AGL = Above Ground Level **MSL = Mean Sea Level Note: If the UAS has even one characteristic of the next level, it is classified in that level.				

Table 1: UAVs Classification according to the US Department of Defense (DoD) [75]

Based on the US classification we can state that Shaheed-136 drones fall into the Group 3 under that classification. In Ukrainian classification drones are classified by the operational capacity, range of the operation and altitude. Let's consider them by functionality and type [79]:

1) Microdrones (“DJI Mavic”): These are small drones that can fit in your hand or pocket. They have a flight range of 1.5 to 10 km.
2) Medium-range drones (“Leleka-100” and “Fury”): These drones have a longer flight range and are used for reconnaissance. Their range of operation varies from 10-50.
3) Operational-tactical drones (PD-2 and Raybird-3): These drones are designed for reconnaissance and operational tasks. The flight range is 50- 600 km.
4) Strategic UAVs: These drones can spend up to 24 hours in the air and fly over 1000 km distances. For instance, “Bober” drone, that weights around 150 kg and can carry a warhead of 20 kg.

Some of the drones that Ukrainian military themselves employ, are classified because of the state of war in the country. Therefore, the information in open sources is very limited about the types of the drones Ukraine produces. But most of the drones that are not classified are operating in the first 3 classes.

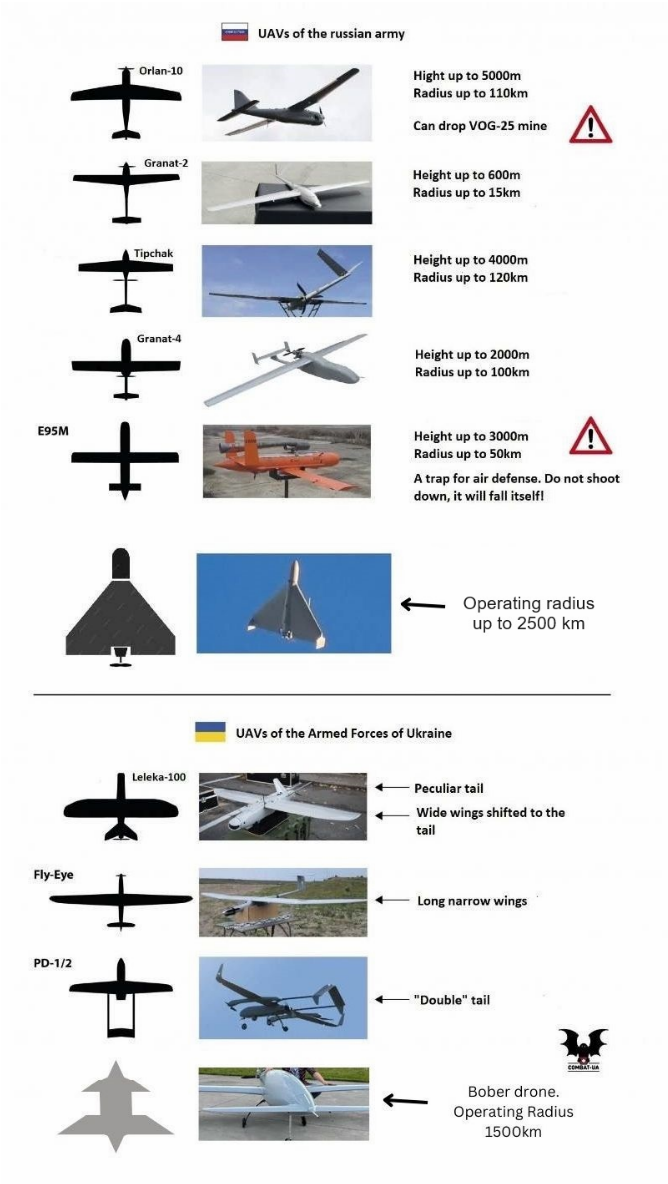

Figure 1. Comparison between Ukrainian and Russian drones in 2022. [72]

Shahed-136, and its smaller Ukrainian counterpart “Bober” are both operating in over 1000 km range and can develop a speed of ~200 km/h [81]. These drones are classified as strategic UAV and are used for in-depth attacks on the enemy’s rear positions [76]. At the beginning of this research in 2022 there were no Ukrainian UAVs that are analogue to the Shaheed-136, therefore we were working on spotting only Shaheed-136/Geran-2 and we didn’t compare the sounds of Ukrainian drones and Russian ones. Moreover, with a very high possibility we would not be allowed to get access to that information in Ukraine as the creation and the models of these drones is top secret information that cannot be shared with anyone. Therefore, this research is out of scope of this thesis. Now, after we understand the classification of our drones, we have to understand the characteristics and the appearance of the drone itself.

The HESA Shahed-136 (Persian: شاهد ۱۳۶, literally "Witness 136"), also known by its Russian designation Geran-2 (Russian: Герань-2, literally "Geranium-2"), is an Iranian-designed loitering munition, also referred to as a kamikaze drone or suicide drone, in the form of an autonomous pusher-propeller drone. [73]

Shaheed-136 characteristics [73]:

1) Mass: 200 kg
2) Length: 3.5 m
3) Wingspan: 2.5 m
4) Maximum warhead weight: 50kg [26]
5) Engine: MD-550 piston engine
6) Operational Range: 2500 km [26]
7) Maximum speed: around 185 km/h (115 mph)
8) Guidance system: GNSS, INS [13]
9) Launch platform: rocket-assisted take-off

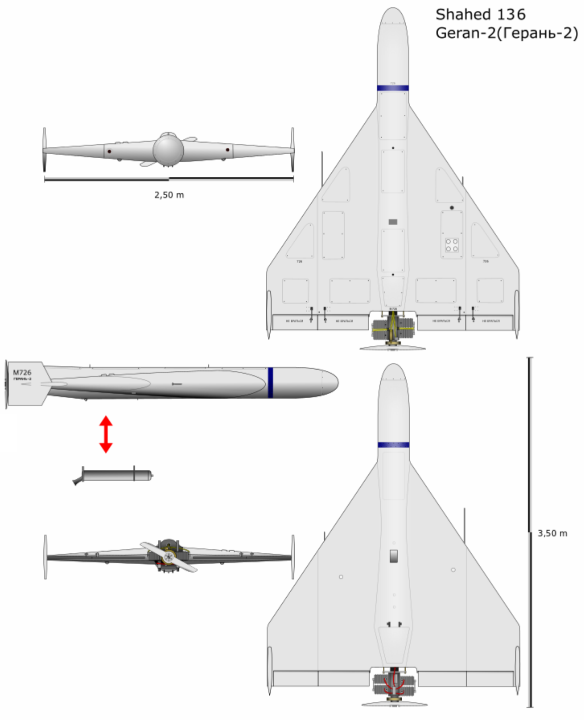

Figure 2. Hesa Shahed-136/Geran 2 [73]

As you can see the size of this drone is rather big, reaching 200 kg of weight in complete assembly. The MD-550 piston engine is a petrol engine that, when operating, produces the special sound of revving that is close to the sound of a motorcycle. Based on that characteristic we were trying to track and classify the sound of the moving drone.

Currently, there are several solutions that are tracking and classifying drone attacks, using both active radar system and acoustic sensors. The Over-The-Horizon radar system is currently widely used to capture strategic types of drones approaching Ukraine, within a range of hundreds to thousands of kilometers [36]. Those are used to just monitor any threats in the air space of the country, as the position of the target is shown very inaccurately at a large distance. Due to their inaccuracy, these systems cannot be used for directing fire, because the speed of the drone is 185 km/h which converts to around 30 m/s and in one minute the drone will move up to 1800 meters away from the place he was before. The smaller range radars are efficient with smaller targets or even the middle-range ones, for example MIM-104 Patriot [88], can track down drones and send missiles to destroy them. However, to cover the whole area of Ukraine the need for those very expensive and deficit systems is enormous [36]. As for acoustic sensors, Ukraine dispatched them in 2024 to help traditional solutions track drones and notify the air defense systems in advance [74]. Those sensors are just microphones that are connected to a GPS tracker and possibly have some simple algorithms for anomaly detection or, perhaps, machine learning models, to help track the drone movement. We cannot know more about those devices as their construction and location are classified, but we can confidently say that these solutions are already used. Next, we will review related literature.

2.1 Related work

The pursuit of advancements in miltech (military technology) has led to increased interest in audio classification systems. These systems are essential for detecting and interpreting sounds of warfare, including gunshots and missile launches. Our research expands upon traditional methodologies, incorporating state-of-the-art deep learning techniques to track down a specific drone.

Signal processing

Before the widespread use of machine learning (ML) and deep learning (DL), sound analysis relied on various signal processing techniques for extracting hand-crafted audio features from the spectrum for interpreting information from audio signals. These traditional methods, grounded in signal processing and statistical analysis, provided foundational approaches for understanding sound in a variety of applications, ranging from speech recognition to environmental sound analysis. This chapter discusses key traditional sound analysis approaches, exploring their principles, applications, and limitations.

A key technique in signal processing is **Fourier Transform**, and, for processing digital audio, particularly its DFT (Discrete Fourier Transform) algorithm and its FFT (Fast Fourier Transform) implementation. FFT converts time-domain signals into frequency components in a computationally efficient way, enabling the analysis of a sound's spectral content. This transformation is essential for identifying frequency bands with significant energy, which is crucial for tasks like speech recognition and soundscape analysis [56].

Short-Time Fourier Transform (STFT) enables to pinpoint the exact time when certain frequencies are heard in a non-stationary signal. A signal is divided into short, overlapping segments and FFT is applied to each segment after amplitude windowing, which reduces artifacts. This provides a **spectrogram**, a time-frequency representation of the signal, essential for detecting temporal variations in sound [2].

Feature extraction is another critical step in traditional sound analysis, where descriptive attributes of audio signals are identified for further analysis. Common methods include **Mel-Frequency Cepstral Coefficients** (MFCCs) and **Linear Predictive Coding** (LPC). MFCCs are particularly effective in speech and speaker recognition tasks, as their computation mimics the human ear's response to varying frequencies, providing a compact sound representation [14].

After extracting spectrum, and optionally dividing spectrum into spectral bands, statistical techniques are applied, such as computing statistical moments, performing Principal Component Analysis (PCA) to reduce dimensionality [10], or Independent Component Analysis (ICA) to separate mixed signals. These techniques are crucial for tasks like noise reduction, echo cancellation, and source separation, where extracting meaningful information from complex audio signals is essential [25].

While these techniques have advanced the field of sound analysis, they face inherent limitations. The primary challenge is the need for hand-crafted features, which require extensive domain knowledge and can lead to suboptimal performance in complex audio environments. Additionally, these approaches struggle to generalize to new or unseen sound patterns, limiting their flexibility and scalability compared to ML-based approaches [10].

In conclusion, traditional sound analysis techniques have laid the groundwork for understanding and interpreting audio signals. Despite their limitations, they remain relevant, particularly in applications where computational simplicity and interpretability are paramount. However, the advent of ML and DL has shifted the paradigm toward more flexible and robust approaches to sound analysis, paving the way for new innovations and applications in the field.

Machine Learning techniques

Machine learning (ML) has revolutionized the field of sound recognition by enabling the analysis of complex audio data to identify and classify a wide range of sound patterns. These range from environmental noises to the distinct sounds of military equipment. ML models can process and interpret large volumes of data, discerning patterns and anomalies that are often key indicators of specific events or threats. This capability is especially crucial in military settings where distinguishing between different types of sounds, even amidst noisy or chaotic environments, is essential for effective surveillance and reconnaissance [48].

In recent studies, researchers from the Hellenic Army Academy and Bolton University have conducted research to classify using a neural network 4 different models of the aircraft based on their spectral centroid and signal bandwidth [5]. To achieve that they studied 4 different types of military aircraft and recorded the sound of their movement. They have achieved 90% accuracy when classifying these 4 types of aircraft. Another study highlights the use of convolutional neural networks in enhancing the accuracy of UAV detection. They have also achieved 96.7% accuracy in identifying drones' sounds [4].

Besides that, there were very promising studies that have focused on developing ML models for sound scene classification, audio tagging, and event detection, employing sophisticated algorithms to process and analyze audio data efficiently. These models can handle a wide range of sound analysis tasks, from environmental sound detection to speech recognition, offering enhanced capabilities for military applications such as surveillance, threat detection, and situational awareness [55,22].

In this thesis, we are using boosted trees, which are a staple algorithm used in many scenarios. XGBoost was used for radar target recognition, which is a task of identifying individual targets by analyzing echo signals received by radar [30]. As concerning application to audio, musical instrument classification is amenable to XGBoost [65]. It is a versatile algorithm that can work well on almost any tasks: early detection of Parkinson's disease [16], phishing URL recognition [28].

Furthermore, distributed analytics for audio sensing applications illustrate how ML can be leveraged at the network's edge, allowing for real-time processing of audio data on devices with limited computational resources. This approach is particularly relevant for military operations, where speed and efficiency in data processing are critical, and connectivity to central servers may be constrained or undesirable for security reasons [64].

Deep Learning

Deep learning models have significantly enhanced the capability to analyze audio signals. These models process audio data represented as sequences of frames, vectors, or tensors, allowing for comprehensive analysis across various dimensions [48].

CNNs have become a mainstay in audio classification due to their ability to work with images, considering that any audio can be represented as an image using a spectrogram. Depending on the nature of the input (spectral features or raw waveform), CNNs use either 1-d temporal or 2-d time-frequency convolutions. Their architecture, consisting of convolutional layers followed by pooling layers, is adept at extracting and down sampling feature maps from audio signals. The optimal architecture for a CNN in audio processing is determined experimentally, with considerations for the task's requirements and data availability [31].

Recurrent Neural Networks (RNNs), particularly effective in modeling sequences, compute outputs by considering both the current input and the previous hidden state. This attribute makes them ideal for capturing temporal dependencies in audio data. Variations like Long Short-Term Memory (LSTM) networks have addressed issues of vanishing and exploding gradients commonly encountered in traditional RNNs. LSTMs utilize gating mechanisms to control the flow of information, making them particularly suitable for audio signal processing where temporal context is crucial [58].

Sequence-to-Sequence Models: In audio processing tasks, sequence-to-sequence models transduce input sequences into output sequences directly. This approach is increasingly being adopted in complex audio processing tasks like automatic speech recognition, where traditional systems involve multiple, separately trained components. Deep learning-based sequence-to-sequence models simplify this by training a single system to map input audio signals to target sequences directly [59].

Audio Analysis in Miltech

Audio analysis in military technology, particularly for threat detection and situational awareness, has significantly evolved from traditional warfare methods, emphasizing the acoustic battlespace. This shift underlines sound's role not only in detection but also as a strategic tool in warfare. The development and advancement of acoustic technology, such as sonar and Long-Range Acoustic Devices (LRADs), reflects this change, offering broad applications across various military branches.

Initially centered on basic detection, sound analysis in military technology now involves complex signal processing techniques capable of deciphering nuanced acoustic signatures in challenging environments. For instance, the U.S. Navy's application of sound technology has transitioned from primarily focusing on submarine detection to exploring its potential as a non-lethal weapon and a strategic tool in warfare [32].

The concept of acoustic warfare has gained traction, viewing sound as more than a means of detection. Technologies like sonar and LRADs are employed for various military applications, from anti-submarine warfare to crowd control, offering non-lethal confrontation means. These technologies have been adapted for different military branches, finding unique operational applications [70].

This includes employing sound for intelligence gathering and direct confrontation in modern warfare scenarios. Additionally, advancements in passive acoustic localization methods emphasize sound's importance in detecting, tracking, and characterizing aircraft and their wake. These methods leverage aircraft acoustic emissions, offering an alternative to conventional RADAR and LIDAR systems [60].

3 Data

In this thesis, our task is to detect a specific model of drone. The presence of other aircraft, missiles, and smaller drones is possible during model inference. We needed a dataset that would contain Shahed-136 drone sound, and other similar sounds, in a battlefield soundscape. We decided to assemble our own dataset.

3.1 Data collection in the field

Together with our partners from Ukrainian military, we have recorded sounds of field explosions and gunshots. The resulting dataset consisted of 10 audios, each 3-5 minutes long. The number of files is very small, but when training it is split into 10 second fragments, which results in ~1500 chunks of audio. We have also conducted some real-life audio recording sessions of Shaheed-136 sound. Sounds were extracted from the videos of flying drones, or just simply recorded in the field during the night shift of the air defense unit. We could only do 3 such collections, because of the danger of the data collection in the field. The Shaheed sound is not audible from the beginning to the end of the recording. The recordings were labeled with Audino, to mark the beginning and end of the sound of Shaheed. We separated the recordings into those that are without any background sound and those that are mixed with other sounds, like gunshots or explosions.

3.2 AudioSet

As the dataset collected in the field was very small, we used publicly available data to increase the size of training data. The best dataset that we found was Google AudioSet [83]. It contains more than 528 classes of different sounds. It is a pivotal resource in audio classification, also for miltech applications. In total, AudioSet has over 2 million 10-second audio clips, categorically diverse, ranging from environmental sounds to human activities. This variety is crucial for training deep learning models to recognize a wide range of audio signatures. In the context of miltech, AudioSet provides specific categories pertinent to aerial military activities. The sound samples of various aircraft types can help models to distinguish between Shahed-136 sound and other aerial vehicles.
From the AudioSet, we took the data from these 3 classes:

1. Aircraft
2. Helicopter
3. Explosion

For the aircraft class, we used data from other similar classes as well, such as airplane, fixed-wing aircraft, etc. This data is used as examples of negative class, that Shahed is easy to confuse with.

3.3 YouTube Scraping

There were no sounds of Shahed drone in Audioset, and we only managed to collect a few in the field. We used YouTube scraping to gather some more data. Scraping is a process of extracting data from the internet, in our case, audio from YouTube videos. The extracted audio is then manually annotated and curated to ensure accurate categorization. To label the data we have used Audino software to collaboratively and quickly label the data. YouTube scraping is essential for enhancing the dataset's diversity, enabling the training of more robust models that can recognize a broad spectrum of sounds in various operational environments. We found 12 publicly available videos with the sound of Shaheed-136 on them. Then after labeling we extracted the chunks and filtered the ones that had a clean sound of the drone, and the ones that had other sounds on top of it. We have done that to make sure our model catches the actual sound of the drone because often the sounds are interrupted with explosions and gunshots.

3.4 Data Preprocessing and Augmentation

After collection we have created a pipeline that extracts features from the audio data. The features were specific to each approach we tried, and we will describe them in later sections. In each approach, we used data augmentation for Shaheed-136 sounds, because despite our best efforts we did not have enough recordings of its sound. For the augmentation, we used "audiomentations" [84] library and have used random transforms for each of the data chunks. Figure 3 shows 4 augmentations that we used:

- Reversing audio sound
- Masking a random segment of the audio completely (making it silent)
- Adding Gaussian noise with a very small loudness (amplitude)
- Making an audio longer or shorter (time stretch). It is important to note that this was done without changing the pitch (the height of the sound).

```
transforms = Compose([
    aud.Reverse(0.7),
    aud.TimeMask(min_band_part=0.1, max_band_part=0.15, fade=True, p=1.0),
    aud.AddGaussianNoise(min_amplitude=0.0001, max_amplitude=0.0002, p=0.7),
    aud.TimeStretch(min_rate=0.8, max_rate=1.25, p=0.5),
], p=0.9)
```

Figure 3: Transformations that were used on the audio using audiomentations library to increase the database size

3.5 Data Distribution

Combining all the sources, we obtained 856 audio files. The last addition was a **Null** class containing various other sounds that might occur in a soundscape in war area, like, motor revving, birds singing, gunshots, explosions, etc. The distribution of the data over classes is shown in Table 2.

Class name	Audio count
Explosion	125
Aircraft	225

Fixed-wing aircraft, airplane	115
Propeller, airscrew	105
Helicopter	100
Shaheed	81
Null	105

Table 2. Data distribution

Classes **aircraft, fixed-wing aircraft, airplane, propeller, airscrew** are all the sounds close to the aircraft, therefore they were included in the data to understand which one will work the best and how the model will work with classifying various similar sounds. Later, we reduced the classes only to Aircraft, Helicopter, Explosion and Shaheed, as those are the most difficult to distinguish between. The explosions were added to help remove those sounds from the detection and also to notify about the explosion nearby the device.

4 Drone sound recognition

The main challenge is distinguishing Shaheed-136 drone sound from other similar sounds: helicopter, aircraft, aircrew, fixed-winged airplane. In this chapter, we will describe two deep learning based multi-class classification models we tried for this scenario.

4.1 CNN + LSTM approach

The motivation behind this approach was that the sound of a drone might be similar to a lot of other sounds (motorcycle, airplane, explosion in a distance), but it has a very different dynamic. A motorcycle moves on the ground and the sound is likely to fade much faster than a drone, that approaches and leaves on a much larger time scale. Therefore, combining CNN for feature extraction and LSTM for long range patterns might be a good idea.

Using deep learning for audio classification is a challenging task, typically requiring a large amount of labeled data and a powerful deep learning model to achieve good performance. The most challenging part about it is the data preprocessing and feature extraction, because the audio samples need to be preprocessed to ensure that they are in a format that can be fed into a deep learning model. This may include resampling the audio to a specific **sampling rate**, and converting the audio to some representation that can be used as input to the model [23].

The most popular way of computing such representation is calculating a spectrogram of the audio. A spectrogram encodes the frequency spectrum of an audio signal. It is a 2-dimensional image where the x-axis represents time, the y-axis represents frequency, and the color or intensity of each point represents the amplitude of the frequency component at that point in time. In other words, a spectrum describes “how much” of a sound we have at a particular moment in the audio [23].

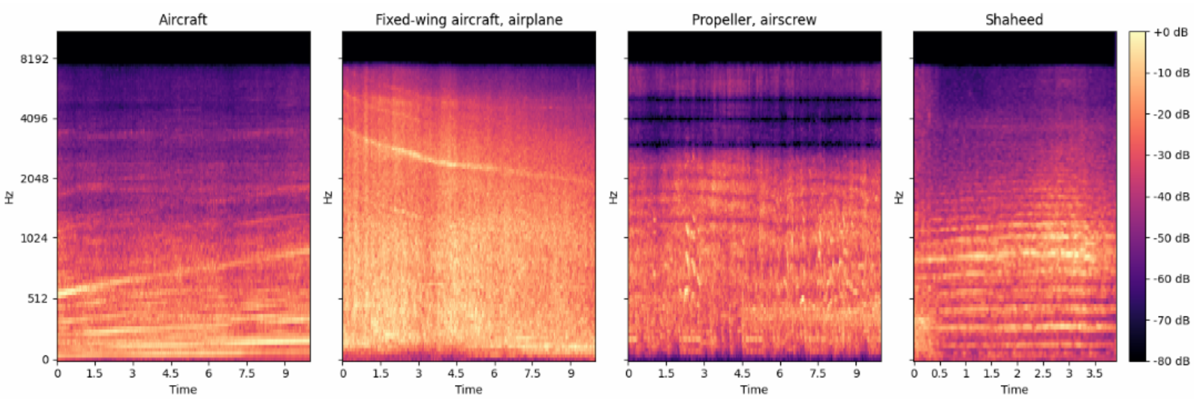

Figure 5: Example of spectrogram computed for our classes.

Dealing with raw spectrograms would imply building an extremely deep neural network to account for different patterns in such high dimensional data. Instead, it is more practical to use **a pretrained model** which would extract **embeddings** from this data and build your architecture from these lower dimensional inputs. Here is the high-level diagram of the model architecture we employ:

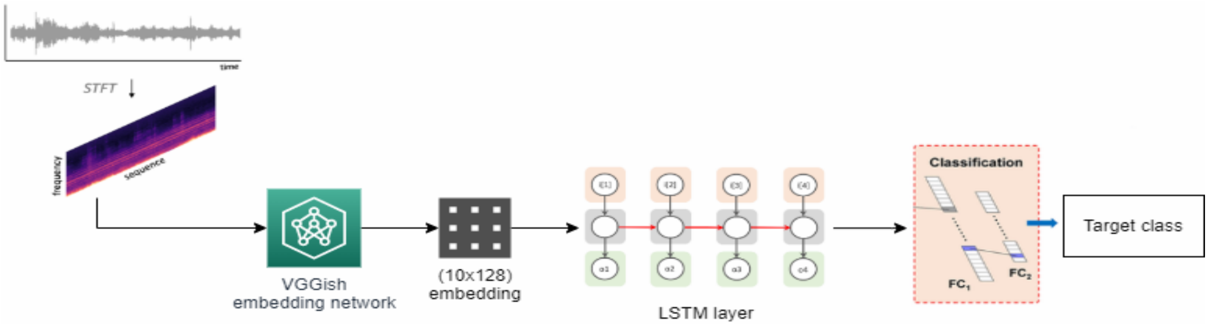

Figure 5: Architecture of the model

To extract embeddings, we use a pre-trained model called VGGish [85], which is a **CNN** architecture from Google. VGGish is trained on the entire Audioset dataset and is able to extract rich embeddings that represent the audio in a more compact and meaningful way given the respective **melspectrogram [56]**. Essentially, VGGish is not trained to handle temporal information in the data it processes. Instead, for the N-seconds audio, it outputs a 128-dimensional vector for each 1-second fragment in an audio, so the output dimensions are (N, 128). In order to train our model we are taking 8-10 second audio chunks, with a sampling rate of 16000, hop length of 160 and window size 10. We have used these classes from our dataset: "Aircraft" - 225 audio chunks, "Fixed-wing aircraft, airplane" - 115, "Propeller, airscrew" - 105, "Shaheed" - 81. We have trained our model for around 30 epochs.

To introduce the sequential nature of the data, we stack a 2-layered LSTM with three fully connected layers on top of the VGGish [85]. The recurrent layer processes the sequence of embeddings one by one and passes the final hidden output to the fully connected layers to perform the classification. The classification logits are activated using the `softmax` function to determine a relative probability for each of the 4 classes.

Evaluation

The CNN+LSTM is trained for 30 epochs with cross entropy loss. We use PyTorch as the main deep learning framework for conducting training experiments. During training, we augment the audio samples with transformations from the audiomentations library, as mentioned before.
Following modern research papers that work on the Audioset, we use the same evaluation metric called F1 score [86]. The model achieves **F1 score = 0.74** with multiclass classification and **0.96** on binary calssification on the validation set. Here is our confusion matrix and precision with recall for the Shaheed and other objects:

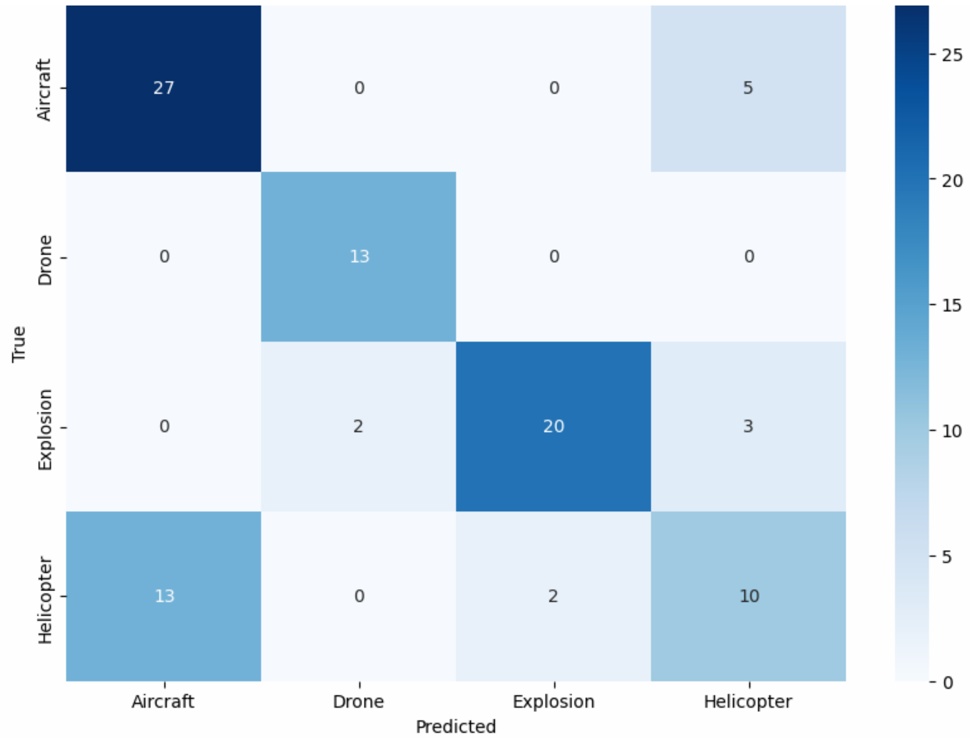

Figure 6: Confusion matrix for multiclass classification, 4 classes. Shaheed recordings are in "Drone" class.

Metric	Value
Average precision	0.77
F1 score	0.75
Precision	0.75
Recall	0.76

Table 4. Average precision, F1 score, precision of the model and recall for multi-class classification.

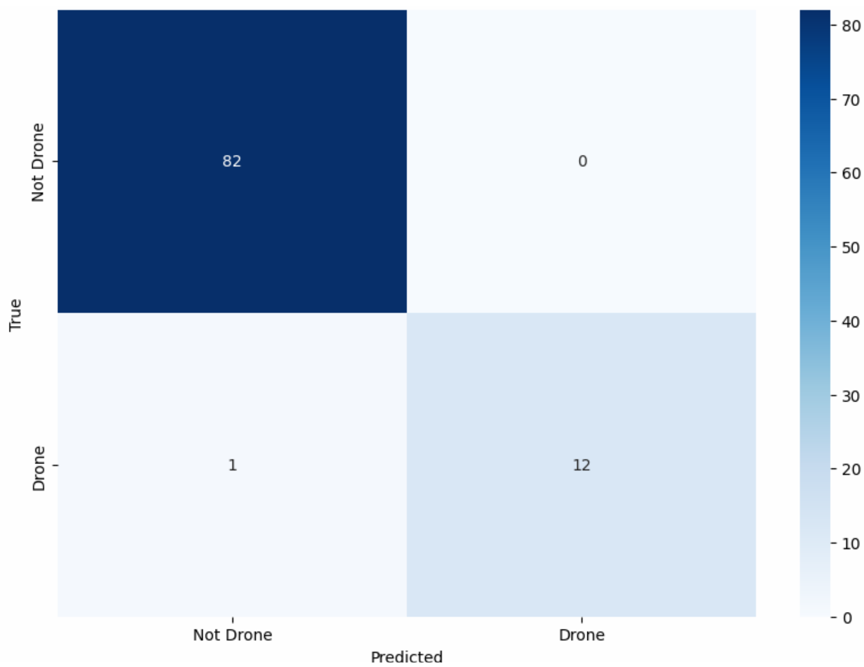

Figure 7: Confusion matrix for Shaheed detection vs other aerial vehicles. There is only false negative, where the system failed to detect Shaheed sound.

Metric	Value
Recall (Drone)	0.92
F1 Score (Drone)	0.96
Precision (Drone)	1.0

Table 5. Binary recall, precision and f1 score for LSTM model for Shaheed vs everything else.

Because we are mostly interested in classifying Shaheed from all the rest of the classes, please note that 13 out of 14 instances of Shaheed are classified correctly.

4.2 CNN approach

VGGish CNN model takes a very short 1-second time window as input, also, the segments that we feed into the model are not that long anyway, as we mostly have recordings with Shaheed already flying overhead in our dataset, and not the ones where it would approach from far away.
We decided to try a state-of-the-art sound recognition CNN model EfficientAT that accepts 10-second-long audio segments.
EfficientAT is an open-source model developed by Google. The process is overall very similar to the one described for CNN+LSTM approach, the dataset is the same and augmentation is applied to the data.

Principles Behind EfficientAT

The core innovation of EfficientAT lies in its ability to distill complex patterns learned by transformer models into CNNs, resulting in a significant reduction in model complexity without sacrificing performance. This process involves training CNNs using the softened outputs (logits) of an ensemble of transformer models as labels, thereby transferring the "knowledge" from the teachers (transformers) to the student (CNN). The ensemble approach, aggregating outputs from multiple transformer models, ensures a rich and diverse set of patterns for CNN to learn from. The choice of MobileNetV3 as the student model aligns with the objectives of achieving high efficiency and performance, given its design to be lightweight yet powerful. [24]
Also, it opens a possibility to do transfer learning on the pretrained models that have already been trained on similar classes of audio.

Model Name	Config	Params (Millions)	MACs (Billions)	Performance (mAP)
dymn04_as	width_mult=0.4	1.97	0.12	45.0
dymn10_as	width_mult=1.0	10.57	0.58	47.7

dymn20_as	width_mult=2.0	40.02	2.2	49.1
mn04_as	width_mult=0.4	0.983	0.11	43.2
mn05_as	width_mult=0.5	1.43	0.16	44.3
mn10_as	width_mult=1.0	4.88	0.54	47.1
mn20_as	width_mult=2.0	17.91	2.06	47.8
mn30_as	width_mult=3.0	39.09	4.55	48.2
mn40_as	width_mult=4.0	68.43	8.03	48.4
mn40_as_ext	width_mult=4.0 ,extended training (300 epochs)	68.43	8.03	48.7
mn40_as_no_im_pre	width_mult=4.0 , no ImageNet pre-training	68.43	8.03	48.3
mn10_as_hop_15	width_mult=1.0	4.88	0.36	46.3
mn10_as_hop_20	width_mult=1.0	4.88	0.27	45.6
mn10_as_hop_25	width_mult=1.0	4.88	0.22	44.7
mn10_as_mels_40	width_mult=1.0	4.88	0.21	45.3
mn10_as_mels_64	width_mult=1.0	4.88	0.27	46.1
mn10_as_mels_256	width_mult=1.0	4.88	1.08	47.4
MN Ensemble	width_mult=4.0 , 9 Models	615.87	72.27	49.8

Table 6. Pretrained models and their parameters [82]

MobileNetV3 is a highly efficient architecture for convolutional neural networks, optimized for performance on mobile and embedded devices. It represents the culmination of automated architecture search techniques and traditional network design principles. The architecture introduces two significant innovations: a modified version of the squeeze-and-excitation block and the use of the h-swish activation function, which are both designed to improve efficiency without compromising accuracy. The squeeze-and-excitation block recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels, thereby allowing the network to focus on more informative features. The h-swish activation function, on the other hand, is a piecewise linear approximation of the swish function that reduces computational complexity while maintaining non-linearity essential for deep learning models [54]. Table 5 shows various model variations available from efficientAT GitHub.

Training and evaluation setup

To adapt EfficientAT models to custom dataset, the straightforward process of transfer learning is applied. Pre-trained models offer a solid foundation for transfer learning, allowing for fine-tuning on specific audio tagging tasks.

Model Selection: We have chosen a pre-trained EfficientAT model that best matches our resource constraints and performance needs. The models vary in complexity, providing options for deployment from edge devices to more capable computing environments. In our case we chose “mn10_as”, “mn20_as” and “mn30_as” that have the following parameters as presented in table 5. We have used that architecture because it’s efficient enough and does not occupy a lot of space on the hard drive, which is beneficial for use on embedded devices.

Data Preparation: Our audio data is processed into a compatible format (Mel spectrograms stored in .hdf format) ensuring consistency with the pre-training conditions of the selected mode. All the sounds from the LSTM + CNN approach, as well as those sounds that might be heard during the operation of the model, were assembled in the “Background” class.

Fine-Tuning: Only the final layer of the model was retrained, including modifying the output layer to match the number of target classes in the dataset. The fine-tuning process involves launching the trained model retraining on the new dataset.

We have tried 2 approaches, one with freezing some of the model layers, and the second one to simply run the model with all the layers employed. The previous model was trained on the FSDK50 dataset that consisted of over 200 various classes from Audioset. For the transfer learning we have frozen 8 first layers and run training on the 3 chosen models for 100 epochs. We did the same thing with the models with all the layers except the last frozen, and again ran them for 100 epochs. After doing that we have also reduced Explosion class in training and validation sets to be at 100 and 40 audios respectively to match the overall distribution of other classes. Then, we removed the “Background” class completely from the training loop as it was having too many sounds that are related to the main 5 classes and repeatedly confused our model.

Evaluation

Here we will present our results with EfficienceAT models pretrained on FSDK50 (mn10, mn20 and mn30) for 100 epochs with 8 frozen layers and same learning rate 5e-5. The weight decay was applied during the training and the model was validated on the separate chunks of the data.
We achieved 82% of AUC for all classes. The mAP reached is about 60% from all 5 classes, meaning there is a lot of misclassification, and we need to analyze where it happens.

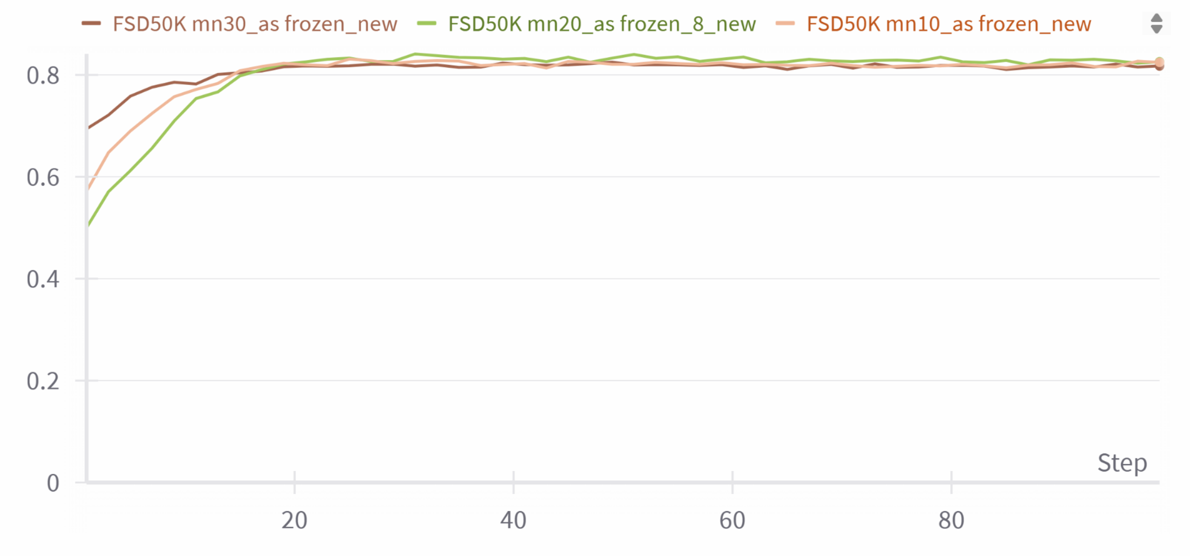

Figure 8: AUC for dataset for all 3 model sizes

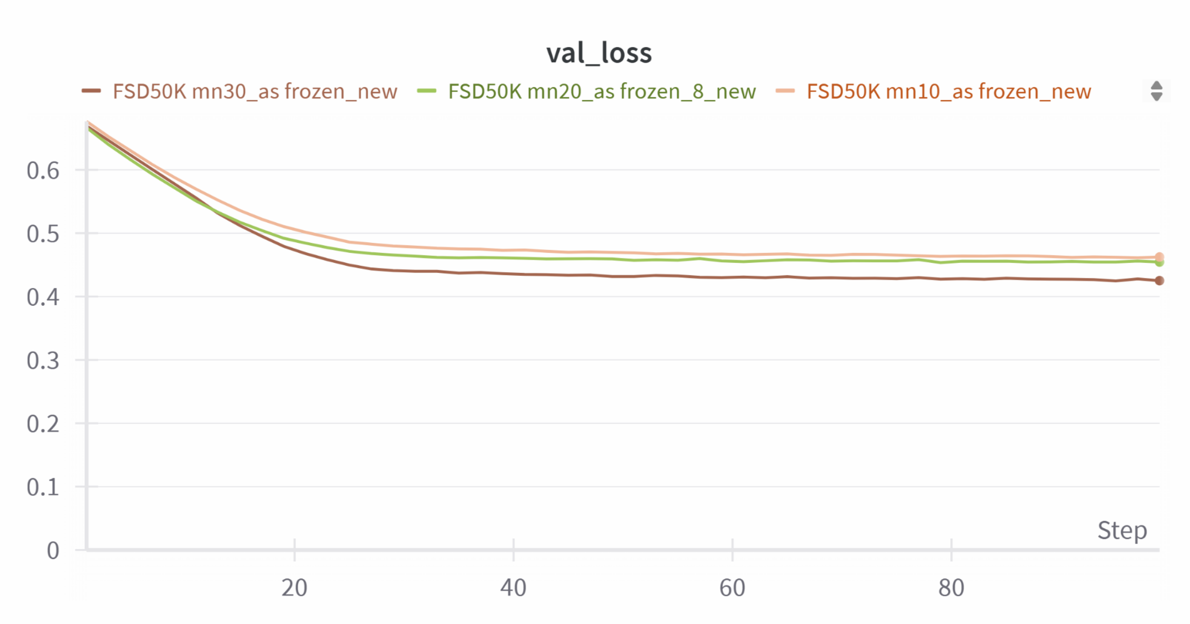

Figure 9: Validation loss for all 3 models.

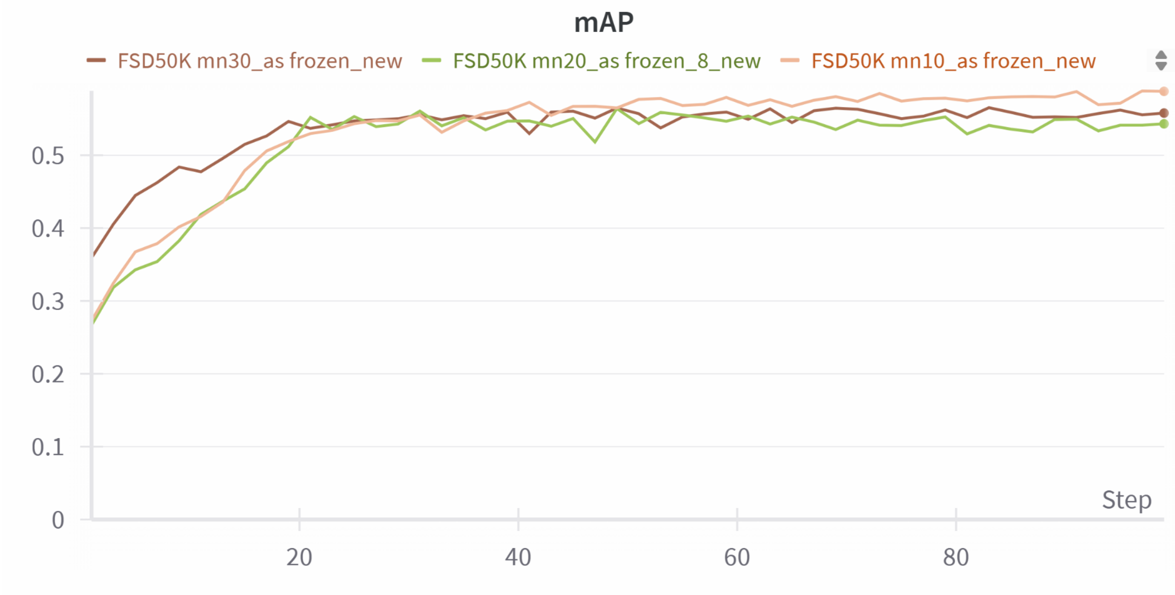

Figure 10: mAP for 3 models. All 3 models in 30 epochs have reached approximately the same mAP. But mn10_as generalize slightly better on all 5 classes

As an additional experiment, we have tried training a model for 30 epochs without freezing the layers.

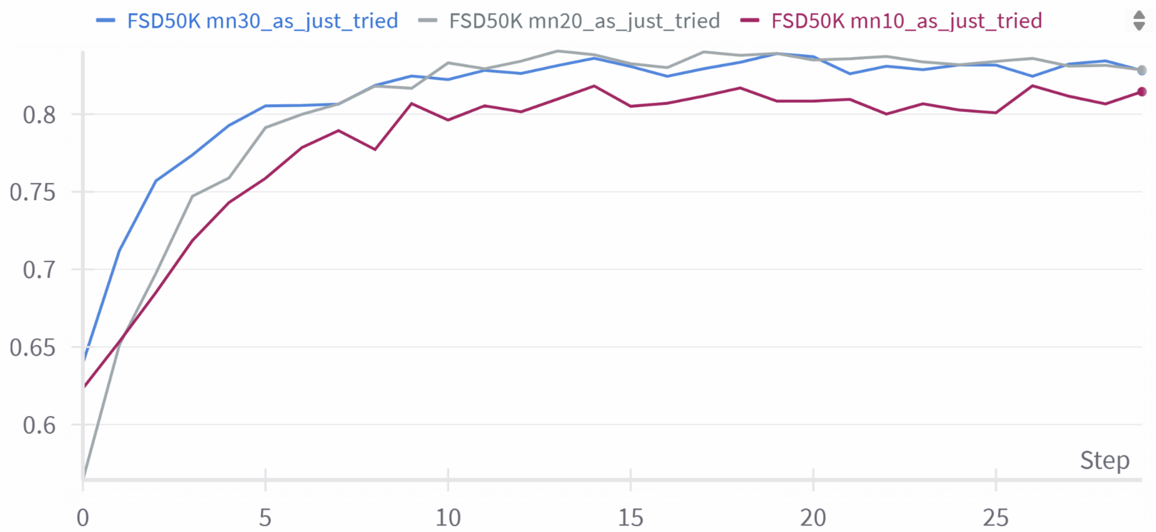

Figure 11: AUC for 3 models with fully activated layers

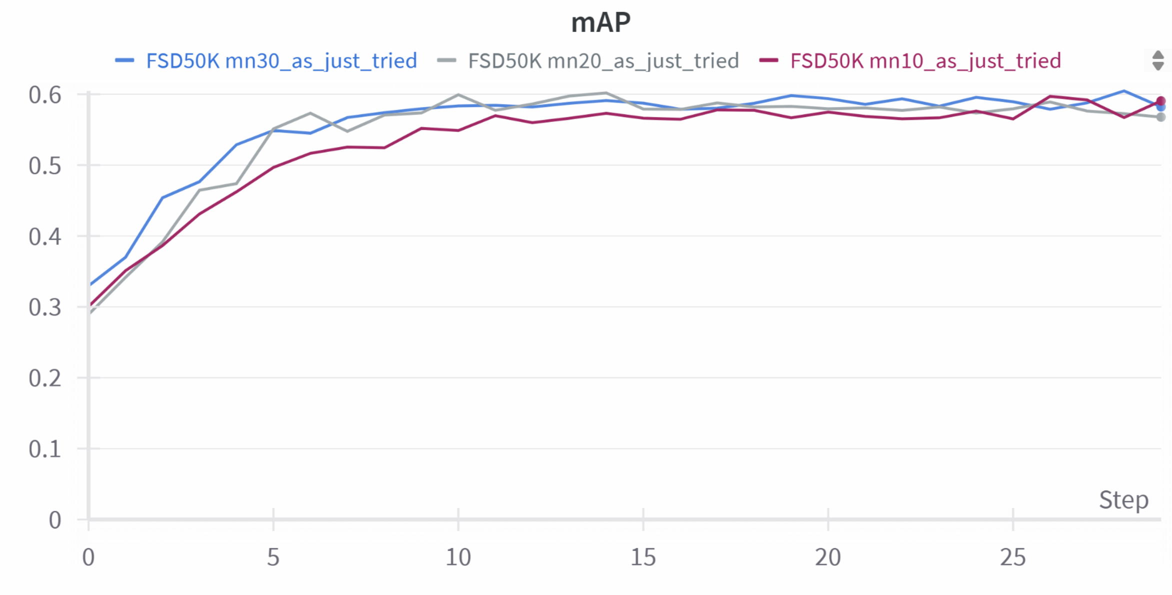

Figure 12: mAP for 3 models with fully activated layers

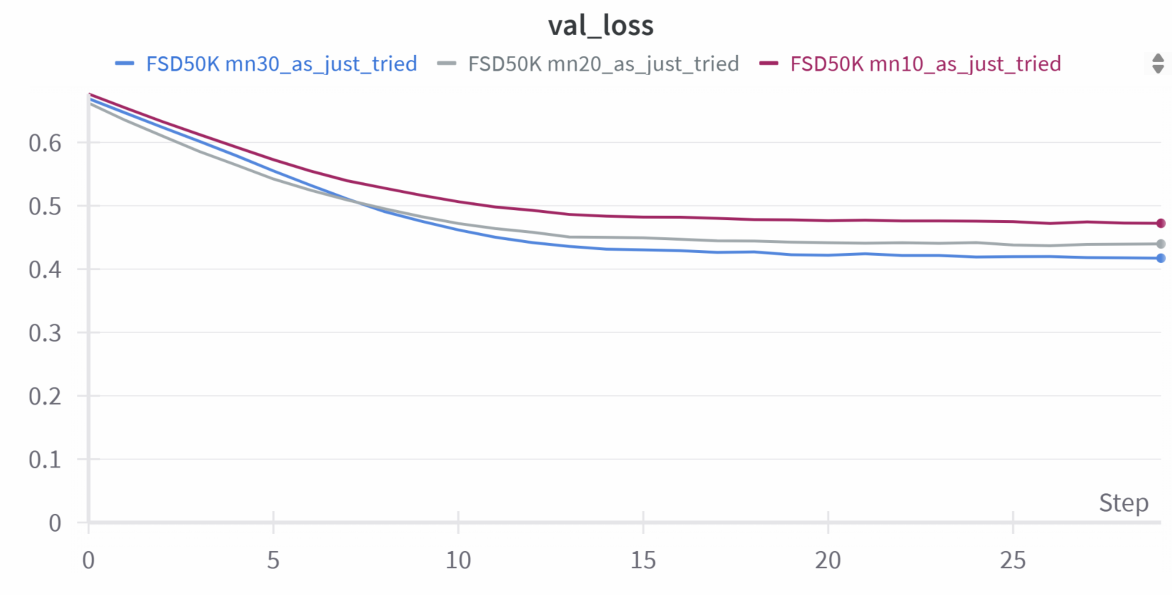

Figure 13: Validation loss for 3 models with fully activated layers

Figure 13 shows a confusion matrix for the mn10_with the same dataset of 5 classes:

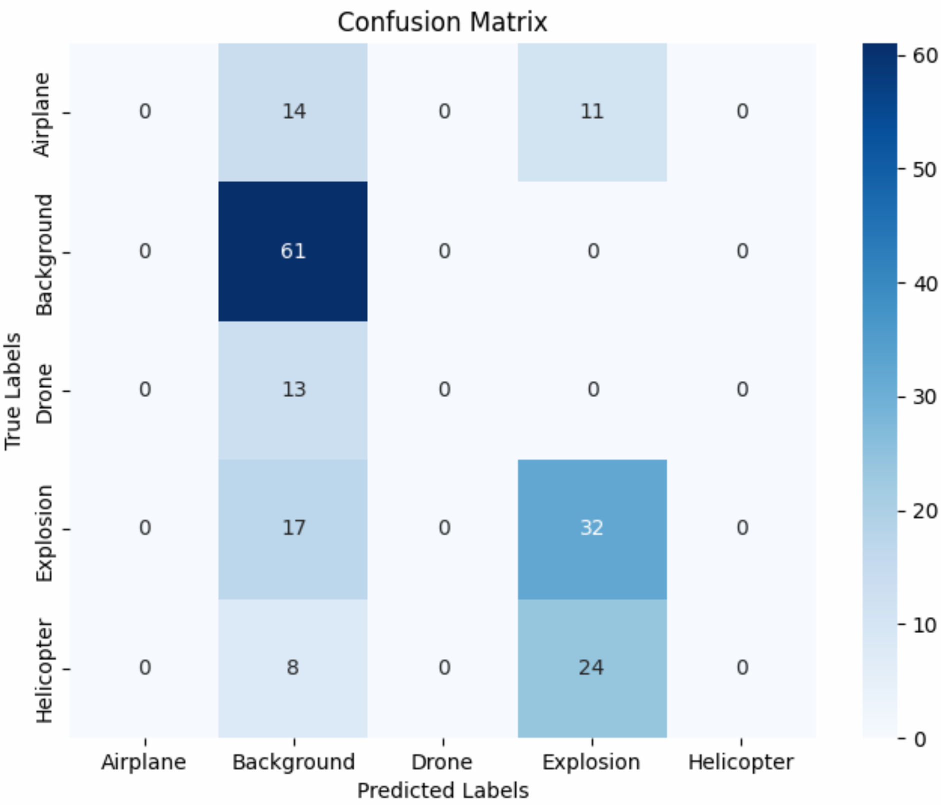

Figure 14: confusion matrix on mn10_as fine-tuned on all 5 classes

Unfortunately, the drone class is not detected correctly, and all the classes are predicted as background. We probably need to clean the "Background" class that probably has some of the classes related to our main ones, which confuses the model. For the next attempt, we have removed the "Background" class and trained our model with 8 frozen layers and 4 classes. Also, we have balanced the dataset more by removing some of the explosion audios, since this class was overrepresented.
As a result, our model were able to converge better and give meaningful outputs:

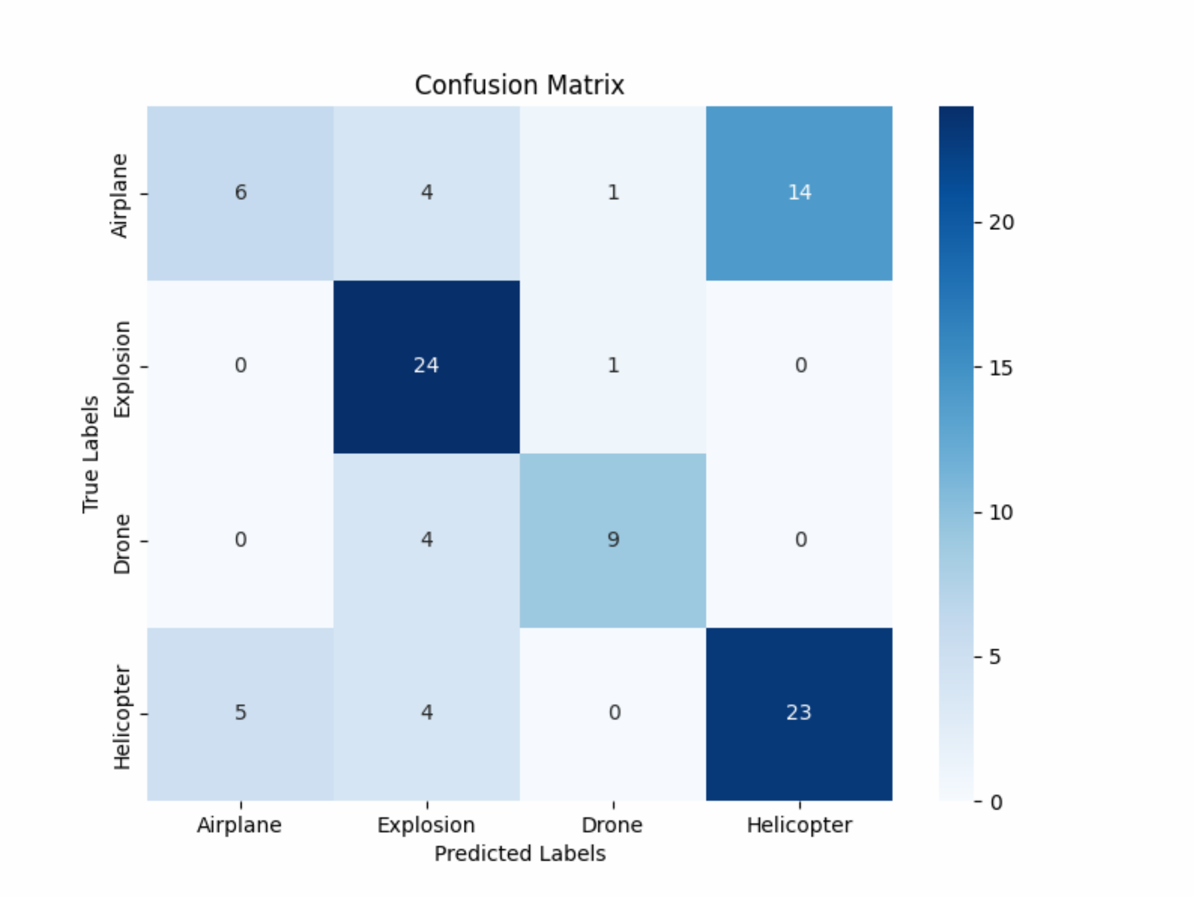

Figure 16: confusion matrix on mn10_as fine-tuned on new data

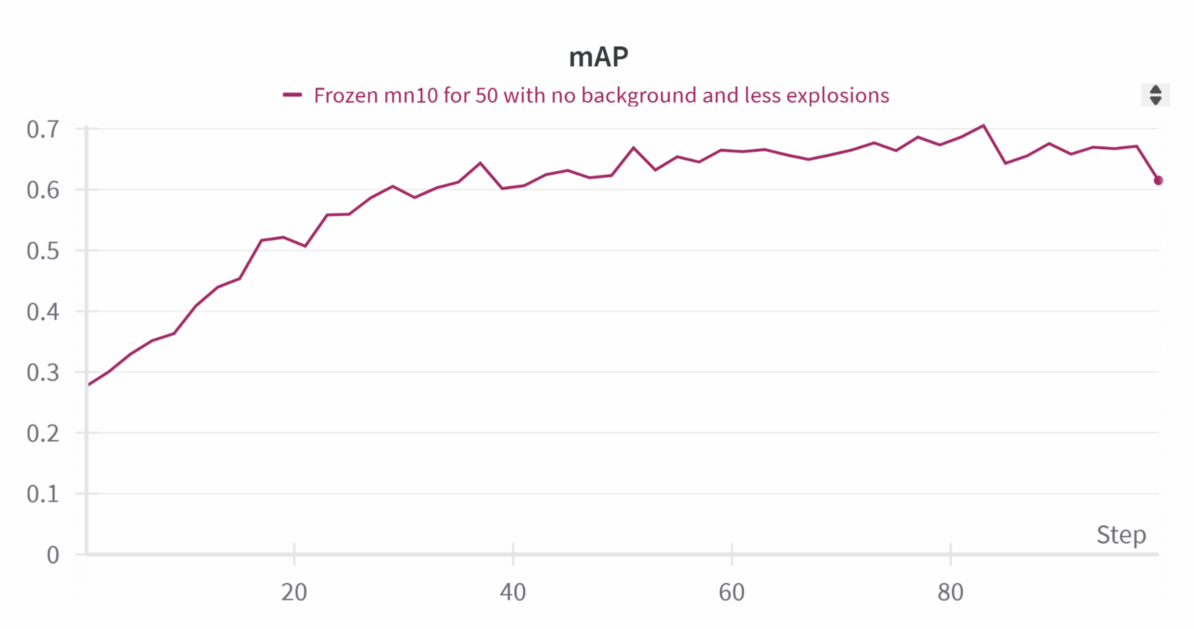

Figure 17: mAP on mn10_as fine-tuned on new data

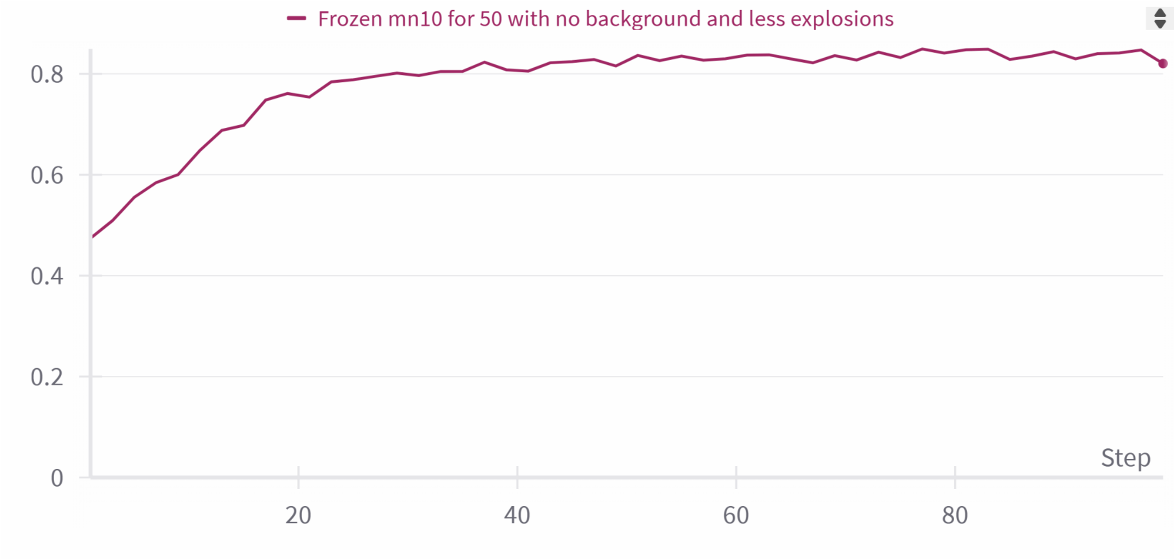

Figure 18: AUC on mn10_as fine-tuned on new data

Therefore, we were able to achieve a mAP of around 0.7 on the validation set with the AUC at 0.82. Also, our confusion matrix clearly shows that the model is able to differentiate between the drone and other types of main classes that can appear during device operation.

Binary classification

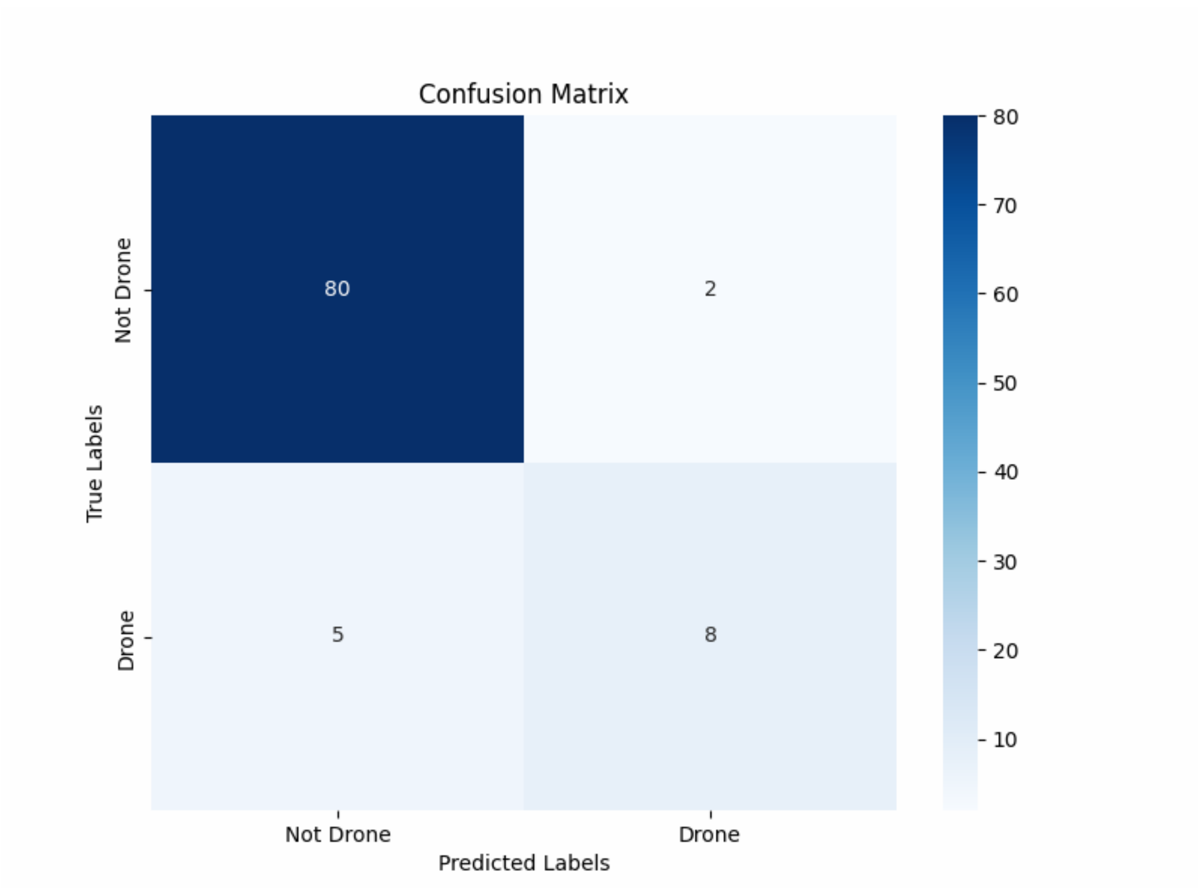

Figure 18: confusion matrix on mn10_as fine-tuned on new data binary classification

To evaluate how the model the binary classification was conducted to differentiate between drones and non-drones. The results show a precision of 0.80, indicating that 80.0% of the instances predicted as drones were correctly identified, and a recall of 0.615, suggesting that 61.5% of actual drone instances were successfully detected, leading to an overall F1 score of 0.696.

Metric	Value
Recall (Drone)	0.615
F1 Score (Drone)	0.696
Precision (Drone)	0.8

Table 8. Binary recall, precision and f1 score for CNN model

5 Discussion

In this thesis we developed a system to distinguish drone sounds from other aerial vehicles with some success. We could achieve an f1-score of 0.96 when distinguishing drone sounds from other aerial vehicles using CNN+LSTM architecture. Distinguishing aerial vehicles between themselves was harder, with an f1-score of 0.74. At the time when the system was developed, only Russian military were using drones in Ukrainian sky, so we did not include any sounds from Ukrainian drones, e.g., Bober. It would be useful to include other drones into the dataset to see whether we can distinguish specifically enemy drones. When expanding the dataset, we should also work on adding battlefield soundscapes, and recording approaching drone sounds, not just overhead drone sounds.

It would be very useful to be able to localize the drone. During the early stages of the project, we worked on localization system based on time difference of arrival. This method works well for explosive sounds that can be precisely timed. There were however several limitations. First, the microphones have to be located about 50 meters apart for triangulation, which proved to be impossible in mobile and dangerous battlefield conditions, where the device also needs to be mobile and fit on an armed vehicle. Secondly, the drone sound is not explosive in nature and cannot be precisely timed. Phase difference of arrival does not work well on sounds that are consist of noise and variable frequency, although in future we might think about adapting it for our purposes.
We also started working on a hardware prototype for the system, but it was abandoned.

Some other problems that we leave for future work are the following:

1) During our development we had various problems with data augmentation and the dataset itself. Since our dataset is scarce, we must use augmentations to generate a bigger dataset for our approach. However, our data augmentations have sometimes led to overfitting of our model.
2) The “Background” class had too many samples, which led to the model overtraining on background noise. The one solution here is to remove the similarly sounding data from the dataset and fill it with other classes that could be heard in the device usage area but are not related to each other.
3) The scarcity of data might lead to the overall decrease of the model performance therefore there is a huge need in collecting as many samples as possible, which is complicated due to the war in Ukraine right now.
4) We were able to fine tune our EfficientAT CNN model to work with our custom data and achieve mAP of 0.71 and binary f1 of 0.69. Meaning we can differentiate between the drones and other airborne targets, but we still must do more work to make the production ready model.

Conclusion

In this thesis we tried to provide a solution for drone attack detection, as fast as possible to help our fellow citizens and protect Ukraine.

We tried implementing a system that would be able to detect a specific drone sound (Shahed-136). We assembled a dataset of Shahed sound and other airplane sounds and implemented two systems, one CNN+LSTM based one and another based on transfer learning on CNN system. As a result, CNN+LSTM model yielded better results achieving stunning f1 score of 0.96 for binary classification problem, while EfficientAT CNN model results in a 0.696 on the same dataset. Thus, LSTM model outperforms CNN on the scarce dataset and can correctly classify drone over other targets in most of the cases.

References

[1] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

[2] Allen, J. B. (1977). "Short Time Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform." IEEE Transactions on Acoustics, Speech, and Signal Processing. doi:10.1109/TASSP.1977.1162950.

[3] Altayeva, A., Omarov, N., Tileubay, S., Zhaksylyk, A., Bazhikov, K., & Kambarov, D. (2023). Convolutional LSTM Network for Real-Time Impulsive Sound Detection and Classification in Urban Environments. International Journal of Advanced Computer Science and Applications.

[4] Anwar, M. Z., Kaleem, Z., & Jamalipour, A. (2019). Machine Learning Inspired Sound-Based Amateur Drone Detection for Public Safety Applications. IEEE Journals & Magazine

[5] Barbarosou, M., Paraskevas, I., & Ahmed, A. (2016). "Military aircrafts' classification based on their sound signature." Aircraft Engineering and Aerospace Technology, 88, 66-72. doi:10.1108/AEAT-04-2014-0040.

[6] Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M., Gannot, S., & Deledalle, C. (2019). Machine learning in acoustics: Theory and applications. The Journal of the Acoustical Society of America, 146(5), 3590.

[7] Brandstein, M., & Ward, D. (1997). Microphone Arrays: Signal Processing Techniques and Applications. Springer.

[8] Camastra, F., Vinciarelli, A., & Yu, J. (2009). Machine Learning for Audio, Image and Video Analysis. J. Electronic Imaging.

[9] Carter, G. C., Knapp, C. H., & Nuttall, A. H. (1973). Estimation of the Magnitude-Squared Coherence Function Via Overlapped Fast Fourier Transform Processing. IEEE Transactions on Audio and Electroacoustics, AU-21(4), 337-344.

[10] Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi:10.1145/2939672.2939785.

[11] Conner, M., Gral, L., Adams, K., Hunger, D., Strelow, R., & Neuwirth, A. (2022). Music Generation Using an LSTM. ArXiv, abs/2203.12105.

[12] "CS231n Convolutional Neural Networks for Visual Recognition". cs231n.github.io. Retrieved 9 March 2023.

[13] Dangwal, Ashish (19 October 2022). "Russia Has 'Upgraded' Iranian Shahed-136 Kamikaze Drones to Boost Its Lethality & Accuracy -- Military Experts". eurasiantimes.com. Retrieved 13 September 2023.

[14] Davis, S., & Mermelstein, P. (1980). "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences." IEEE Transactions on Acoustics, Speech, and Signal Processing. doi:10.1109/TASSP.1980.1163420.

[15] Deep Convolution Neural Network sharing for the multi-label images classification, Machine Learning with Applications, Volume 10, 2022, 100422,ISSN 2666-8270,https://doi.org/10.1016/j.mlwa.2022.100422.

[16] Francisco, D. (2023). Parkinson's Disease Detection using XGBoost and Machine Learning.

[17] Ge, R. (2023). XGBoost-Based Human Activity Recognition Algorithm using Wearable Smart Devices. Applied and Computational Engineering.
[18] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
[19] Guo, G., & Li, S. (2003). "Content-based audio classification and retrieval by support vector machines." IEEE transactions on neural networks. doi:10.1109/TNN.2002.806626.
[20] Gupta, K. D., Nigam, R., Sharma, D., & Dhurandher, S. K. (2022). LSTM-Based Energy-Efficient Wireless Communication With Reconfigurable Intelligent Surfaces. IEEE Transactions on Green Communications and Networking, 6, 704-712.
[21] Hakkani-Tür, D. Z., et al. (2016). "Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM."
[22] Heittola, T., Çakir, E., & Virtanen, T. (2018). The Machine Learning Approach for Analysis of Sound Scenes and Events.
[23] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Zhang, Z. (2017). CNN architectures for large-scale audio classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
[24] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., & Adam, H. (2019). Searching for MobileNetV3. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
[25] Hyvärinen, A., & Oja, E. (2000). "Independent Component Analysis: Algorithms and Applications." Neural Networks. doi:10.1016/S0893-6080(00)00026-5.
[26] "IRGC confirms specs for Shahed-136 attack UAV". Janes Information Services. 17 May 2023. Archived from the original on 28 May 2023.
[27] J. Pan, J. Li, P. Hu and Q. Bao, "Experimental Results of Passive Bistatic Radar Using Phased Array Radar as Illuminator," 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2019, pp. 1725-1728, doi: 10.1109/ITAIC.2019.8785843.
[28] Jia, Q., Guo, X., Zhang, M., Liu, M., Tian, X., Jin, X., & Ma, D. (2023). Phishing URL recognition based on ON-LSTM attention mechanism and XGBoost model. 2023 5th International Conference on Electronics and Communication, Network and Computer Technology (ECNCT), 159-163. https://doi.org/10.1109/ECNCT59757.2023.10280927. Liu, Y., Yihods, 17(3), 261-272.
[29] Jiang, W., Cai, Z., Luo, M., & Yu, Z. (2011). A simple microphone array for source direction and distance estimation. 2011 6th IEEE Conference on Industrial Electronics and Applications, 1002-1005. https://doi.org/10.1109/ICIEA.2011.5975733.
[30] Jiang, Y., Cheng, D., Qi, W., Song, R., Yu, C., & Liu, S. (2023). Radar Target Recognition of Individuals Based on XGBoost. 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), 3, 1110-1113. https://doi.org/10.1109/ICIBA56860.2023.10165343.
[31] Kong, Q., et al. (2019). "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing. doi:10.1109/TASLP.2020.3030497.
[32] Koza, John R.; Bennett, Forrest H.; Andre, David; Keane, Martin A. (1996). "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits

Using Genetic Programming". Artificial Intelligence in Design '96. Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151f–170. doi:10.1007/978-94-009-0279-4_9. ISBN 978-94-010-6610-5.
[33] L. M. Encarnacao, R. J. Barton and D. Zeltzer, "Interactive exploration of the underwater sonar space," MTS/IEEE Oceans 2001. An Ocean Odyssey. Conference Proceedings (IEEE Cat. No.01CH37295), Honolulu, HI, USA, 2001, pp. 1945-1952 vol.3, doi: 10.1109/OCEANS.2001.968144.
[34] LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015). "Deep Learning". Nature. 521 (7553): 436–444. Bibcode:2015Natur.521..436L. doi:10.1038/nature14539. PMID 26017442. S2CID 3074096.
[35] Lundberg, S. M., & Lee, S. (2017b). A unified approach to interpreting model predictions. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
[36] Malyasov, Dylan (2020). "Ukrainian Army receives new radar system that can spot combat drones." Defence Blog. Retrieved April 9, 2024.
[37] Markel, J. D., & Gray, A. H. Jr. (1976). "Linear Prediction of Speech." Springer-Verlag.
[38] Marouani, H., & Dagenais, M. (2005). Comparing high resolution timestamps in computer clusters. Canadian Conference on Electrical and Computer Engineering, 2005., 400-403. https://doi.org/10.1109/CCECE.2005.1556956.
[39] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In Proceedings of the 14th Python in Science Conference.
[40] Mekruksavanich, S., Jantawong, P., & Jitpattanakul, A. (2022). LSTM-XGB: A New Deep Learning Model for Human Activity Recognition based on LSTM and XGBoost. 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON).
[41] O. Ennasr and X. Tan, "Time-Difference-of-Arrival (TDOA)-Based Distributed Target Localization by A Robotic Network," in IEEE Transactions on Control of Network Systems, vol. 7, no. 3, pp. 1416-1427, Sept. 2020, doi: 10.1109/TCNS.2020.2979864
[42] Oppenheim, A. V., & Schafer, R. W. (2009). Discrete-Time Signal Processing. Pearson.
[43] Pasquali, V., D'Alessandro, G., Gualtieri, R., & Leccese, F. (2017). A new data logger based on Raspberry-Pi for Arctic Notostraca locomotion investigations. Measurement, 110, 249-256.
[44] Pan S. J. and Yang, Q. "A Survey on Transfer Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.
[45] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
[46] Pham-Quoc, C. (2023). Using LSTM Network for Predicting Radio Mobile Networks' Behaviors.
[47] Punuri, S. B., Kuanar, S. K., Kolhar, M. S., Mishra, T. K., Alameen, A., Mohapatra, H., & Mishra, S. (2023). Efficient Net-XGBoost: An Implementation for Facial Emotion Recognition Using Transfer Learning. Mathematics.

[48] Purwins, H., et al. (2019). "Deep Learning for Audio Signal Processing." IEEE Journal of Selected Topics in Signal Processing. doi:10.1109/JSTSP.2019.2908700., “Applications”
[49] Quinn, Joanne (2020). Dive into deep learning: tools for engagement. Thousand Oaks, California. p. 551. ISBN 978-1-5443-6137-6. Archived from the original on January 10, 2023. Retrieved January 10, 2023.
[50] Remigius Meier, Armin Rigo, and Thomas R. Gross. 2016. Parallel virtual machines with RPython. In Proceedings of the 12th Symposium on Dynamic Languages (DLS 2016). Association for Computing Machinery, New York, NY, USA, 48–59. https://doi.org/10.1145/2989225.2989233n
[51] Rong, F. (2016). "Audio Classification Method Based on Machine Learning." 2016 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 81-84. doi:10.1109/ICITBS.2016.98
[52] Ruitenberg, Rudy (2023). "Germany to supply Ukraine with IRIS-T systems in a $1.4 billion package." C4ISRNET. Retrieved November 24, 2023.
[53] Schmid, F., Koutini, K., & Widmer, G. (2023). Efficient Large-Scale Audio Tagging Via Transformer-To-CNN Knowledge Distillation. arXiv:2211.04772v3 [cs.SD].
[54] Schmid, F., Koutini, K., & Widmer, G. (2023). Efficient Large-Scale Audio Tagging Via Transformer-To-CNN Knowledge Distillation. arXiv:2211.04772v3 [cs.SD].
[55] Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. Applied Acoustics, 158, 107020.
[56] Smith, J. O. (2007). "Spectral Audio Signal Processing.
[57] Solemane Coulibaly, Bernard Kamsu-Foguem, Dantouma Kamissoko, Daouda Traore,
[58] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks." ArXiv.
[59] Tang, Y., et al. (2016). "Sequence-to-Sequence Model with Attention for Time Series Classification." 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). doi:10.1109/ICDMW.2016.0078.
[60] Tenney, S.M., Mays, B., Hillis, D.B., Tran Luu, D., Houser, J., & Reiff, C.G. (2004). Acoustic Mortar Localization System - Results from OIF.
[61] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
[62] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., ... & van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Met
[63] W. Feng, F. Lu, T. Pu, X. Chen, Y. Guo and Q. Zhang, "Intelligent Target Detection for Airborne STAP Radar Based on Space-Time Image Feature and YOLO," 2023 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), ZHENGZHOU, China, 2023, pp. 1-6, doi: 10.1109/ICSPCC59353.2023.10400319.
[64] Wood, D., Wang, S., Salonidis, T., Conway-Jones, D., Ko, B., & White, G. (2018). Distributed analytics for audio sensing applications. , 10635. https://doi.org/10.1117/12.2306096.
[65] Y., Zhu, Q., & Cui, W. (2022). Musical Instrument Recognition by XGBoost Combining Feature Fusion. ArXiv, abs/2206.00901.

[66] Yang, C., Gan, X., Peng, A., & Yuan, X. (2023). ResNet Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments. Sustainability.
[67] Zaman, K., et al. (2023). "A Survey of Audio Classification Using Deep Learning." IEEE Access. doi:10.1109/ACCESS.2023.3318015.
[68] Zhu, Q., Bao, Q., Hu, P., & Chen, Z. (2018). Experimental Study of Aircraft Detection by PBR Exploiting Uncooperative Radar as Illuminator. DEStech Transactions on Computer Science and Engineering.
[69] https://en.wikipedia.org/wiki/Spectrogram, Spectrogram - Wikipedia
[70] https://apac.genasys.com/wp-content/uploads/LRAD-Product-Guide-Final-PRINT.pdf
[71] https://numpy.org/
[72] https://en.defence-ua.com/weapon_and_tech/how_to_distinguish_between_russian_and_ukrainian_uavs_in_the_sky_photo_comparison-3180.html
[73] https://en.wikipedia.org/wiki/HESA_Shahed_136
[74] https://www.twz.com/land/thousands-of-networked-microphones-are-tracking-drones-in-ukraine
[75] https://www.e-education.psu.edu/geog892/node/5
[76] https://en.wikipedia.org/wiki/Unmanned_aerial_vehicle
[77] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
[78] https://en.wikipedia.org/wiki/Speed_of_sound
[79] https://24tv.ua/yaki-droni-vikoristovuye-ukrayini-viyni-yaki-buvayut-bezpilotniki_n2359466
[80] https://en.wikipedia.org/wiki/Time_of_arrival#:~:text=Time%20difference%20of%20arrival%20%28TDOA%29%20is%20the%20difference,TOA%20as%20signals%20travel%20with%20a%20known%20velocity.
[81] https://en.wikipedia.org/wiki/Bober_(drone)
[82] https://github.com/fschmid56/EfficientAT/tree/main
[83] https://research.google.com/audioset/, Google Dataset
[84] https://github.com/iver56/audiomentations
[85] https://github.com/tensorflow/models/blob/master/research/audioset/vggish/README.md
[86] https://blog.paperspace.com/mean-average-precision/
[87] https://en.wikipedia.org/wiki/Cartesian_coordinate_system
[88] https://en.wikipedia.org/wiki/MIM-104_Patriot

Non-exclusive licence to reproduce the thesis and make the thesis public

I, __________Roman Karpenko____________________________________,
(*author's name*)

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Drone sound recognition and tracking system using machine learning and acoustic localization

,
(*title of thesis*)

supervised by ____________Anna Aljanaki ________________________________.
(*supervisor's name*)

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Roman Karpenko
16/05/2024