2022 Sixth IEEE International Conference on Robotic Computing (IRC) | 978-1-6654-7260-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IRC55401.2022.00009

UAV Payload Detection Using Deep Learning and Data Augmentation

Ilmun Ku
Artificial Intelligence Convergence
Hankuk University of Foreign Studies
Seoul, South Korea
mun90505@hufs.ac.kr

Seungyeon Roh
Computer Science and Engineering
Konkuk University
Seoul, South Korea
shtmddus99@konkuk.ac.kr

Gyeongyeong Kim
Computer Science and Engineering
Sunmoon University
Asan, South Korea
kky57389@sunmoon.ac.kr

Charles Taylor
Computer and Information Technology
Purdue University
West Lafayette, United States
taylo869@purdue.edu

Yaqin Wang
Computer and Information Technology
Purdue University
West Lafayette, United States
wang4070@purdue.edu

Eric T Matson
Computer and Information Technology
Purdue University
West Lafayette, United States
ematson@purdue.edu

***Abstract*—In recent years, the technology behind Unmanned Aerial Vehicles (UAVs) has continually advanced. However, with these developments, malicious activities employing UAVs have also been on the rise. Within this study, Deep Learning (DL) algorithms are utilized to detect and classify UAVs transporting payload based on the sound they release. In order to exercise DL algorithms on a set of data, a sufficient amount of audio data is necessary to obtain a more reliable result. So UAV sound recordings have been collected alongside the use of data augmentation to secure a satisfactory sample size for testing purposes. Afterward, a feature-based classification was applied to the groups of audio identifying each UAV's payload (or lack thereof). Lastly, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Convolutional Recurrent Neural Network(CRNN) are utilized in analyzing the final data-set. They are evaluated for their abilities to correctly categorize the unloaded, one payload, and two payload of UAV classes and noise class solely through audio. As a result, MFCC showed the best performance in CNN, RNN, and CRNN, which are 0.9493, 0.8133, and 0.9174 accuracies. Our contribution to this study is that a cost-efficient data collection method was applied by utilizing laptop microphones. Moreover, DL technology was used in UAV payload detection, whereas neural network was used in prior study. Also, the best feature for UAV payload detection with the three DL technologies was found. The limitation of the paper is that only two UAV models and one kind of payload were used to collect data. Diverse UAVs and payload are expected to be used to collect data in future works.**

***Index Terms*—UAVs, payload detection, deep learning**

I. Introduction

Over the years as Unmanned Aerial Vehicle (UAV) technology has continued to advance drones have become much more accessible to the public. While UAV accessibility has been steadily growing, malevolent activities have also become increasingly common. Especially UAVs with payload that can easily be employed to endanger innocent civilians or Government Dignitaries with their airborne contents. These unknown packages could be potentially transporting harmful materials, explosives, etc. through public airspace.

In 2019, a Houthi drone was observed targeting senior Yemeni military officers and even exploded over a military base, killing 5 and wounding 20 according to [1]. In another instance, a UAV was seen carrying a bottle with radioactive material and even landed on the Japanese Prime Minister's living quarters according to [2]. Additionally, in 2018 there was an infamous incident involving Venezuela's President Maduro who was attacked by two small drones carrying explosives [3].

In [4], Machine Learning (ML) algorithms were applied to detect UAVs carrying payload on the basis of the sound that they emit. Feature extraction in addition to a combination of other features is proposed for preparing data for further analysis. In this study, however, these methods will be utilized in order to train and evaluate Deep Learning (DL) models as opposed to ML Algorithms. Although in the past the study had been limited to using only the ML algorithms for detecting UAVs due to the limited processing power available when using a MacBook Air with only 8GB of RAM.

In order to deploy DL algorithms to the classification model, extra resources and additional data are required. Furthermore, data collection has been conducted with a Raspberry Pi in previous research. Although there are more effective methods, with this study there is an aim to develop a method that is convenient and easy to use by anyone. In accordance with [5], data augmentation methods are pivotal for the smooth and continuous improvement of audio classification performance utilizing DL algorithms.

The rest of this paper is organized as follows. Section II inspects the current acoustic identification methods for UAV classification and payload classification. Section III presents data augmentation methods. Furthermore, the methodology for the feature extraction and two different DL models is suggested. Section IV describes our experiments and results. Finally, Section V presents the conclusion and future work.

978-1-6654-7260-9/22/$31.00 ©2022 IEEE
DOI 10.1109/IRC55401.2022.00009

II. Literature Review

A. Audio UAVs Classification

As the threat of spiteful drones has continued to rise, UAVs have gradually become an important issue. Thus, various UAVs and their associated fields of research have evolved to be a field teeming with interesting and new research. Numerous detection and classification methods have been developed, including image classification, audio classification, utilization of radar, and so on. Since audio data is less affected by weather conditions and light when compared to image data, audio UAV classification can realistically be considered a reliable method to detect UAVs.

In the study [6], research about autonomous detection systems for Unmanned Aerial Vehicles (UAVs) based on acoustic signatures has been examined. In order to train DL models, a large amount of drone audio data is essential in order to detect the differences between them. The researchers recorded a UAV audio clip with an iPhone, alongside a clip solely consisting of background noise. After both were recorded and properly formatted the separate clips of audio were split into single-second segments for processing. This method of data processing is dependent upon transforming audio clips into spectrograms for further analysis.

In the study, researchers made use of two different brands of UAVs for the purpose of data collection. Then previously recorded samples of background noise, alongside samples of loaded and unloaded drones were recorded for testing purposes [6]. Both subsets of labeled data were then merged, and CNN, RNN, and Convolutional Recurrent Neural Network (CRNN) were utilized to evaluate and compare each neural network's performance. As a result, CNN outperformed RNN and CRNN in both accuracy and F1 score. Nonetheless, the performance of CRNN had a small difference in proportion to the performance of CNN.

In another study, research utilizing a combination of audio data and visual data has been applied and analyzed for it's effectiveness [7]. Audio and image data-sets were obtained through several sensors placed within the zone that the researchers had marked off for audio collection at various distances. Feature extraction was performed by applying the Mel-frequency Cepstral Coefficient (MFCC) descriptor and Image feature extraction had been implemented through the usage of AlexNet [8]. Afterward, a Support Vector Machine (SVM) algorithm was deployed in the actual analysis of the image and auditory data-sets. Several different kernel methods were tested on the data-set. From the methods tested RBF/Gaussian kernel's analysis of the combined audio and imagery data-set were shown to produce the highest rate of accuracy. This accuracy rate was noticeably higher than the other methods: linear, and polynomial kernels.

According to prior research, UAVs can be identified through ML algorithms based on UAV sound signals. Five features, including MFCC, chroma, Tonnetz, contrast, and Mel were analyzed for their effects on sound data analysis. To better train the ML model, a combination of those features is then exploited to produce more consistent results [8]. The ML model can produce an acceptable result with a relatively small data-set. In this study, features extracted from raw data are sent to the ML model as input, with that raw data the model is then trained, and subsequent model evaluation is then performed upon the data. Certain combinations of features were found to produce better performance than others, so among the combinations analyzed, MFCC and chroma were found to bring about good performance. Although there were certain limitations to this experiment as there were only two UAVs being utilized, and the data-set the researchers made use of was considerably small.

Shi et al [9], feature extraction from UAV audio data-sets was carried out by applying MFCC. The researcher had even employed classification analysis by utilizing Hidden Markov Model (HMM). A collection of Twenty-four MFCCs and a collection of thirty-six MFCCs are procured in order to extract feature vectors. These specific amounts are chosen to demonstrate a correlation between the sample size and the reliability of the results. A classifier based on HMM identifies their two features. As the larger the sample size of sounds the model includes in training, the higher the average recognition rate will rise. Therefore, they have concluded that the method in this study is more effective in distinguishing UAV sounds regardless of whether or not they are located within a noisy environment.

Kartashov et al [10], UAV detection and direction-finding were conducted utilizing Bartlett and Capeon's methods of cross-correlation function. The data-sets that were classified were acoustic emission that was collected via the microphone array. In spectral power density (SPD), the first 10 spectral peaks included 80% of the acoustic signal energy. For this reason, first spectral peaks are used when UAV detecting and recognizing. The data-set consists of numerous kinds of UAV acoustic signatures in different conditions. The best accuracy was achieved only up to 55m/70m for detection/recognition.

B. UAV payload detection

Since drones have become rather widely used around the globe, there are hazards that come with the rise in the popularity of UAVs, namely harmful gas or explosives. There are quite a few papers detailing practical and potential methods of detecting a UAV's payload. In [4], UAV payload identification had been effectuated through the usage of sound signals and a classification system making use of a Machine learning (ML) algorithm. Features were extracted from the audio data-set by using librosa since it is a feature-based classification. 5 features: mfcc, chroma, tonnetz, contrast, mel, and their different combinations were trained by the ML model, several different algorithms were thus utilized including SVM, Neural Network (NN), Gaussian Naive Bayes (GNB), and K-Nearest Neighbors (KNN). The ensuing comparison between the differing combinations and their rates of performance were then put into focus [4]. The dataset used for testing consisted of sound samples involving both loaded drones and unloaded

drones. Their payload method involved hanging a 500ml bottle of water from the base of the drone. The result of the training was that the combination of MFCC, Mel, Chroma, Contrast, and Tonnetz shows higher accuracy than individuals. Among individuals, MFCC and chroma show high performance. Albeit there are caveats, for instance, a smaller set of data, or the usage of only a single payload during audio collection.

In a separate study utilizing YOLOv2, both loaded and unloaded UAV detection utilized a solely image-based dataset [11]. For data collection purposes, a DJI Phantom 2 drone with an attached 100g object for payload was used. To overcome the data shortage, drone images were collected from open source. The performance of loaded and unloaded UAV detection turned out to be 74.97% of mean-average precision. The paper proposed future work involving studying different types of object detectors such as Fast R-CNN, and Mask R-CNN that are believed to improve performance [11]. In [12], the study differentiated loaded and unloaded drone images by classifying them through the use of a residual convolutional neural network.

Similarly, another team of researchers took images of both the loaded UAV and unloaded UAV using the same DJI Phantom 2 drone, this time with a weighted object attached weighing 1000g. After data collection, it was determined that more data was required, so data augmentation was performed on the set of images with different transformations for example rotation, resizing, the addition of Gaussian noise, cropping, and so on. A training of 96% was achieved through the application of ResNet-34. Throughout the research written upon this topic in the past there have been many different methods of detection tested, one study sought to detect hovering micro drones with differing sizes of loaded objects loaded via multistatic radar [13]. Detected Micro-Doppler signatures on radar were noticeably different when spotted without payload, this study tested both a 200g payload and a 500g payload while hovering. Even utilizing classification methods such as Naïve Bayes and discriminant analysis, both of them performed with above 90% accuracy. UAV sound classification was also used in conjunction with the LSTM-CNN architecture in order to classify drones [14]. The multiple labels of collected data were as follows: ‘loaded’, ‘unloaded’, and ‘no drone’. The feature extraction is a data processing process involving the time domain, frequency domain, and MFCC. At last, training on the data-set was proceeded using a stacked bidirectional LSTM-CNN structure. ‘Stacked BiLSTM-CNN structure’ is a combined structure with BiLSTM layers and CNN layers. The result of the combined structure was an accuracy rating of 94.28%, so it is considered to be an efficient model.

III. Methodology

A. Dataset

During the audio data collection two models of drones were utilized the DJI Phantom 4 and the EVO 2 Pro. 1306 samples were collected for the DJI Phantom while 1238 samples were collected for the EVO 2, for a total of 2544 samples, which totals 7.04 hours in length, as shown in Table I. Moreover, 232 noise samples were collected with a total length of 0.64 hours long. Throughout all categories each audio recording is only 10 seconds in length.

	unloaded	1 payload	2 payload	noise
DJI Phantom 4	199	558	549	
EVO 2 Pro	210	507	521	232
total	409	1065	1070	232

TABLE I
Dataset Categories

For each UAV, audio samples were recorded and categorized according to three separate classes; UAV with two payload, UAV with one payload, and unloaded UAV. The two payload utilized within the study weighed 63g and 68g each. Only the 68g payload was used when recording audio under the classification of one payload class. For the single payload set of data 558 samples were collected. 549 samples were collected under the classification of two payload. Each of these were recordings of the DJI Phantom 4. For the EVO 2 Pro 507 single payload samples were collected, and 521 samples with two payload attached were gathered. Since there is only a 63g weight difference between the two one payload classifications, it was deemed necessary to collect more data for these categories than previous tests in an effort to help the algorithm more efficiently classify the data collected.

B. Feature Extraction

Features are defined as values that represent the unique characteristics of sound that could be extracted from audio. The human perception system easily recognizes different types of audio. However, since the model does not have a built-in system that is as inherently astute as humans, so the feature extraction process is required so that the models are better able to understand the sound. In this study, feature extraction is employed through the use of the python library Librosa, providing the extraction method seen in Table II.

Several methods of feature extraction were selected for use with the data-set, Mel-frequency cepstral coefficients (MFCCs), Mel, Chroma, Tonnetz, and contrast. Mel-frequency cepstral coefficients (MFCCs) are values meant to represent the short-term power spectrum of a sound, these values are received through cepstral analysis of a spectrogram . This in turn reflects the relationship between physical frequencies and the human scale of perceivable frequencies. Therefore, MFCCs are commonly used in the analysis of audio data since MFCCs are similar to human perception systems that are not linear but instead much more sensitive to low frequency, and thus rather useful in differentiating between the myriad of various noises heard in daily life, as mentioned in [15] [16]. Mel Spectrogram refers to the conversion of a frequency to Mel scale. Chroma differentiates sound into 12 pitches, C, C#, D, D#, E, F, F#, G, G#, A, A#, B. While Tonnetz makes use of a pictorial representation of the sound and reveal affinities and structures

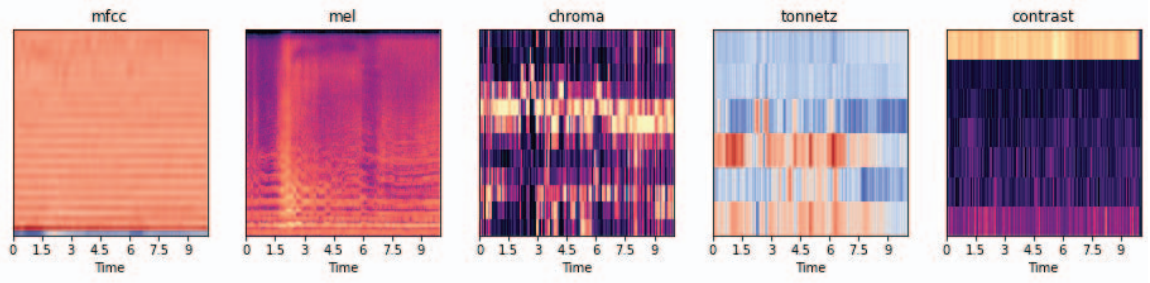

Fig. 1. Extracted Features

Feature	Shape
Chroma stft	12
Chroma cqt	12
Chroma cens	12
Mel	128
MFCC	40
RMS	1
Centroid	1
Bandwidth	1
Contrast	7
Flatness	1
Rolloff	1
Poly shape	2
Tonnetz	6
Zero crossing	1

TABLE II
LIBROSA - FEATURE EXTRACTION

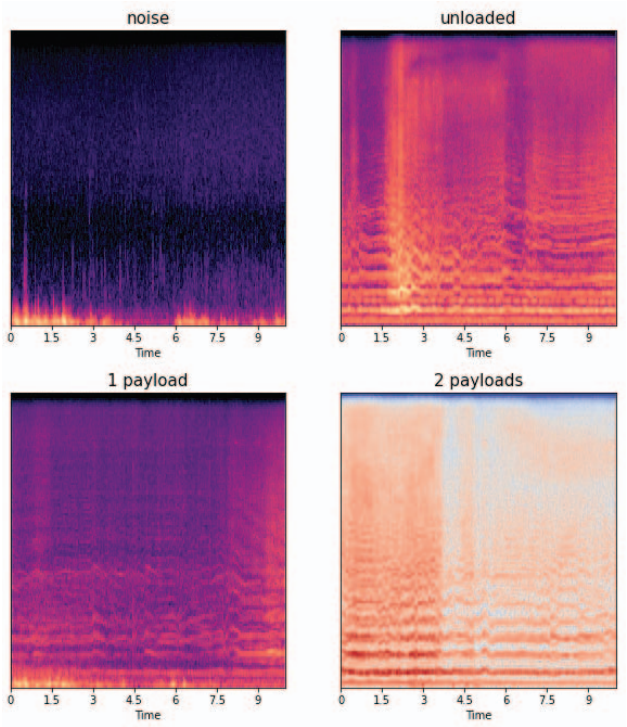

Fig. 2. Class Comparison Among Unloaded, One payload, and Two Payload Using Mel Spectrogram

between notes in order to demonstrate its harmonic relationship. Lastly, Contrast represents the frequency of power which is then measured at each timestamp.

To clearly distinguish the classes, It is better to pre-process the data-set through the use of feature extraction rather than perform analysis on the raw data-set. Additionally, through the use of spectograms, which bring visualization to certain features hidden within audio, the data can be distinguished more easily from one another as compared to unaltered data. Thus the features extracted utilizing librosa are saved as json files and entered as input into the model.

C. Data Augmentation

A large audio data-set is required to properly utilize DL models and solve audio classification problems. When working with sets of data comprised of only text or images more data can be gathered by either randomly altering data or generating new data. However, collecting enough audio samples for effective application of DL is a time-consuming process. If data-sets are insufficient, overfitting and poor generalization may occur in the classification process. Data augmentation technology is one of the solutions to cover the lack of a data-set problem. This method can obtain a new data-set by augmenting the original data-set.

This study uses time-stretching, pitch scaling, time masking, and frequency masking in the augmentation of audio data. Time stretching and Pitch scaling are methods of audio augmentation performed upon raw data, whereas time masking and frequency masking is a method of data augmentation executed upon a spectrogram, treating this process more as a visual problem rather than an audio one.

Raw augmentation is a method of augmenting the audio file itself. It is proven helpful in enhancing accuracy for LSTM-based RNN wielding only raw audio data [17]. There are time-shifting, time stretching, pitch scaling, noise addition, impulse response addition, low/high/pass-band filters, and so

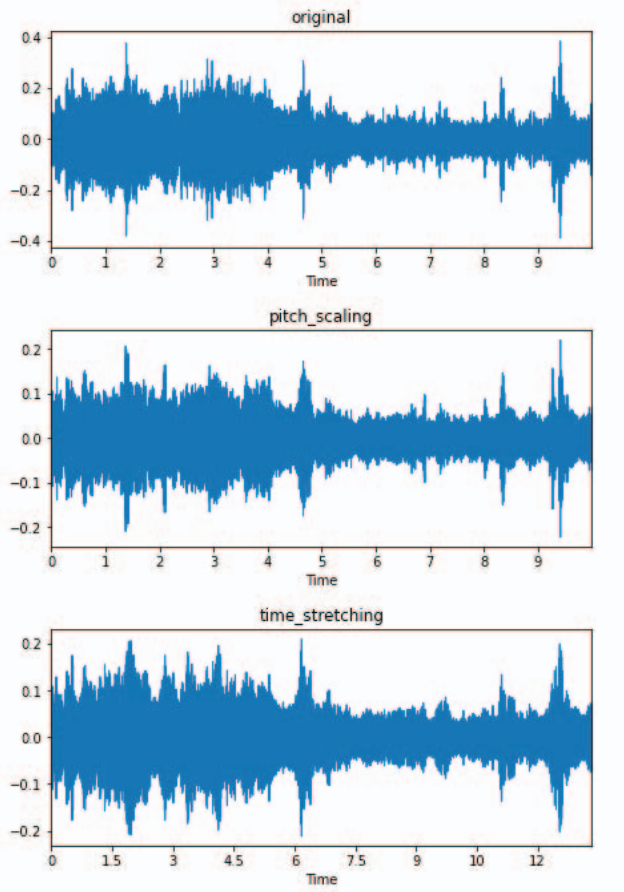

Fig. 3. Raw Data Augmentation Techniques

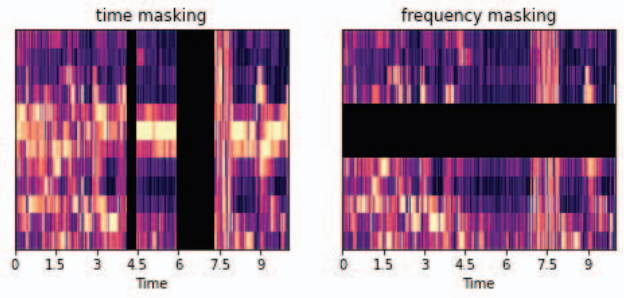

Fig. 4. Spectral Data Augmentation Teqhniques

forth. Time stretching changes the speed without altering pitch or frequency, which can change the speed of the sound, making it either faster or slower. As opposed to time stretching, pitch scaling changes the pitch without changing the speed. For example, C major is changed to D major if a signal is up to 2. It is recognized to be advantageous when exercised in advancing CNN accuracy [18].

Spectrogram augmentation refers to data augmentation performed on a spectrum rather than a raw audio file, this includes operations such as time masking and frequency masking. Time masking obscures a particular part of the spectrum, and those are masked with 0 or minimum value. Blocking off a portion of the time domain, which exists along the x-axis of the spectrum. Frequency masking is the inverse of time masking, instead hiding a part of the spectogram's frequency domain, which exists along the y-axis of the spectrum, by masking a specific part with either a minimum value or 0. These methods are widely known to improve the network performance without the extra arrangement for the network, or hyper-parameter [19]. Helping bolster the networks robustness, prevent overfitting, and fight against deformation by presenting purposely altered/corrupted data.

We obtained 10140 additional data samples by data augmentation and secured 12675 samples in total. Each audio sample is 10 seconds in length. The accuracy of MFCC enhanced from 85% to 94% with CNN and from 87% to 91% with CRNN, although decreased from 85% to 81% with RNN, as shown in Table III and IV.

	Accuracy	Recall	Precision	F-1 Score
CNN	0.8716	0.8743	0.8762	0.8739
RNN	0.8516	0.8534	0.8530	0.8539
CRNN	0.8750	0.8767	0.8771	0.8772

TABLE III
MFCC PERFORMANCE BEFORE DATA AUGMENTATION

	Accuracy	Recall	Precision	F-1 Score
CNN	0.9493	0.9476	0.9502	0.9483
RNN	0.8113	0.8150	0.8185	0.8141
CRNN	0.9174	0.9156	0.9163	0.9159

TABLE IV
MFCC PERFORMANCE AFTER DATA AUGMENTATION

D. Deep Learning Models

Layer	Stride	Out Dim	Activation Function	Kernel Size
Conv2D	(1,1)	16	ReLU	3x3
Average Pooling 2D	(2,2)	16		2x2
Batch Normalization				
Conv2D	(2,2)	32	ReLU	3x3
Max Pooling 2D	(2,2)	32		3x3
Batch Normalization				
Flatten		1		
Dense		32	ReLU	
Dropout(rate = 0.3)				
Dense(Output)		4	Softmax	

TABLE V
CNN ARCHITECTURE

Layer	Out Dim	Activation Function
LSTM	64	tanh
LSTM	64	tanh
Dense	64	ReLU
Dropout(rate = 0.3)		
Dense	4	Softmax

TABLE VI
RNN ARCHITECTURE

Layer	Stride	Out Dim	Activation Function	Kernel Size
Conv2D	(1,1)	16	ReLU	3x3
Average Pooling 2D	(2,2)	16		2x2
Batch Normalization				
LSTM		32	tanh	
Flatten		1		
Dense		32	ReLU	
Dropout(rate = 0.3)				
Dense(Output)		4	Softmax	

TABLE VII
CRNN ARCHITECTURE

CNN has demonstrated itself to be very effective not only in image classification but also in efforts of audio classification generating relatively positive results [20]. CNN operates based upon the extraction of local information from audio signal representations such as MFCCs or Mel spectrogram.

RNN is oftentimes used to process sequential data such as that found in text or sound. However, RNN has its own downsides, namely gradient exploding wherein the derivatives grow exponentially to the point that the gradient becomes immeasurable by the model, and gradient vanishing where the derivatives shrink causing the gradient to exponentially decreases to the point that it eventually vanishes [21]. Through the application of Long Short Term Memory(LSTM) networks the risks of these issues are reduced. As in LSTM layers, self-recurrent weights make the cells in the memory block retain previous information [22] [23].

CRNN is another commonly used neural network architecture that combines features found in both CNN and RNN. Due to this CRNN analyzes both local information and the longer temporal context. The local information is extracted with the help of the CNN, and the longer temporal context is captured by the RNN [24]. In acoustic classification, CRNN has proven to garner noticeable results over the years [6].

IV. EXPERIMENTS AND RESULTS

A. Data Collection

The UAV audio and background noise data were recorded at a remote location located at 62F2+V33 New Richmond, Indiana, USA. Data recording and processing were conducted on a MacBook Pro, with a 2.6 GHz Hexa Core Intel Core i7 and 16GB 2667 MHz DDR4. Audio data was collected with the laptop placed on a 73cm high table, facing toward a field of soybeans, with a forest sitting behind the recording

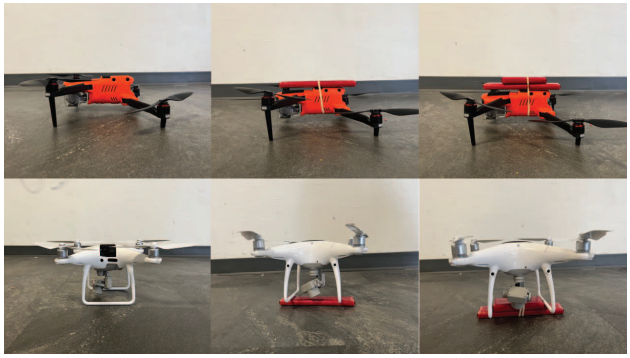

Fig. 5. DJI Phatom4 and EVO 2 Pro unloaded, with one payload, and two payload

environment. Both a DJI Phantom 4 Pro and EVO 2 Pro were employed in the gathering of 7.13 hours of audio samples. Throughout data collection an effort was made to maintain a consistent recording environment. Temperatures were measured from 22 degrees Celsius to 31 degrees Celsius, and wind speeds were between 6.4 km/h and 17.7 km/h. The humidity ranged from 46% to 76%. While the Maximum distance between the laptop and UAV remained 15m long, with a maximum altitude of 10m high.

B. Training Strategies

CNN structure is presented in Table V. The CNN model has two 2D-convolutional layers, two 2D-pooling layers, and one fully connected layer. The first convolutional layer comprises 16 filters with three by three sizes, and its stride is set to 1. The second convolutional layer consists of 32 filters with a three-by-three sizes, with a stride set to 2. Every convolutional layer went through the ReLU activation function, and the average pooling layer with a two-by-two size kernel was assigned right after. After a fully connected layer, the softmax activation function is utilized to predict the classifaction of a UAV payload. Additionally, batch normalization is used to assist in optimizing the training of a model, in conjunction with the dropout method helping to mitigate model overfitting.

The workflow of the CNN classifier is defined as follows. First, graphic form data such as a spectrogram is created by applying a method of feature extraction on a clip of raw audio data. Then, the local information is captured by convolution operation in the convolutional layers. Next the captured local information is fed into a fully connected layer in order to gain the prediction of the model. In the final portion of the CNN classifier, the soft-max function is employed to obtain the target class from the model's prediction.

RNN structure is shown in Table VI. The RNN model has two LSTM layers and two fully connected layers. Each LSTM layer has 64 dimensionalities of output space, and makes use of a hyperbolic tangent as an activation function. The first fully connected layer has ReLU as an activation function, and the last has a softmax function. The dropout rate is set to 0.3.

The Feature data extracted from the audio samples passes

through two LSTM layers. The LSTM layer slightly changes the data using a cell state through either simple multiplication or addition. When data comes out from the LSTM layers, it then flows into a Dense layer with the activation function ReLU. Then finally, the Dense Layer performs a prediction and gives the data a classification.

Our definition of the CRNN framework is presented in Table VII. The CNN block has one 2D-convolutional layer for extracting feature maps and one 2D average pooling layer for the down-sampling of data. The 2D-convolutional layer is comprised of 16 kernels of 3x3 size, and its stride is set to 1x1. The 2D average pooling layer consists of a 2x2 kernel, with a stride set to 2x2. In the RNN block, one LSTM layer is applied with 32 memory units to capture context information. The fully connected layer has 32 hidden neurons to learn trends of data. Lastly, the soft-max function is adopted to obtain the predicted class's probability distribution. Both a batch normalization layer and dropout layer are employed so that the model can prevent overfitting/underfitting problems and achieve stable training. For the CRNN flow, audio signal representations such as MFCCs or Mel spectrogram are first sampled and produced from the raw audio data. Then convolutional layers produce high-level feature maps through the convolution of an audio signal representation. After going through the convolutional layer, LSTM layers are applied to gain the context information. Then, the predicted value from the model is acquired from a fully connected layer. Lastly, as a result of soft-max a classification of data is derived.

The training-related hyperparameters are set equally in the three proposed DL models. The learning rate is initialized at a value of 0.0001, the mini-batch size is 32, and the training epoch is 200. The categorical cross entropy function is made use of in calculating loss at the end of every epoch. Then, the model weights are optimized with the assistance of the Adam algorithm. To prevent overfitting, the early stopping method is employed. Model training is terminated when the calculated loss shows no improvement for ten consecutive epochs. Every DL technology within the study was applied through the use of the Python library Keras.

Feature	Accuracy	Recall	Precision	F-1 Score
MFCC	0.9493	0.9476	0.9502	0.9483
Mel	0.9133	0.9132	0.9151	0.9140
Chroma_stft	0.7883	0.7900	0.7880	0.7872
Contrast	0.7814	0.7795	0.7868	0.7825
Tonnetz	0.5645	0.5602	0.5687	0.5560

TABLE VIII
TEST RESULTS FOR CNN

C. Results

Table VIII presents the performance of the CNN model. While MFCC has proven to produce the best results in accuracy, recall, precision, and F-1 score. Mel also demonstrated reasonable performance under the same criterion. Chroma and contrast revealed very similar results, although slightly under performing in comparison.

Feature	Accuracy	Recall	Precision	F-1 Score
MFCC	0.8113	0.8150	0.8185	0.8141
Mel	0.7247	0.7230	0.7350	0.7253
Chroma_stft	0.5743	0.5669	0.5656	0.5661
Contrast	0.7162	0.7244	0.7218	0.7230
Tonnetz	0.4256	0.4137	0.4789	0.3959

TABLE IX
TEST RESULTS FOR RNN

Feature	Accuracy	Recall	Precision	F-1 Score
MFCC	0.9174	0.9156	0.9163	0.9159
Mel	0.9260	0.9017	0.9069	0.9040
Chroma_stft	0.8182	0.8171	0.8171	0.8168
Contrast	0.8001	0.8049	0.8043	0.8038
Tonnetz	0.7675	0.7734	0.7681	0.7682

TABLE X
TEST RESULTS FOR CRNN

Table IX shows the outcome for the RNN model. MFCC continues to produce an outstanding performance in all four categories. Mel, and Contrast follow behind within a 10% margin. While Chroma and Tonnetz displayed poor performance in all four scores.

Table X exhibits the result for the CRNN model. MFCC is again proven to bring pronounced results in each criteria. Mel also manifested promising results in recall, precision, and F-1 score and even outperformed MFCC in accuracy.

Overall, MFCC showed remarkable and steady performance with all three DL models in accuracy, recall, precision, and F-1 score. Mel also performed well in the same evaluation criteria. Especially in CNN and CRNN, MFCC and mel scores are all above 90%. Chroma showed unstable performance since it presented 57% accuracy in RNN while 78% and 81% accuracy in CNN and CRNN.

V. CONCLUSION AND FUTURE WORKS

In this paper, the automatic classification of UAV payload were implemented through the use of three deep learning techniques; CNN, RNN, CRNN. Deep learning models assorted the audio dataset into four different classes; noise, unloaded, one payload, and two payload. The models not only distinguish whether a UAV is loaded or not, but can even categorize the number of payload. Additional equipment was not required in the experiment as the MacBook pro's built-in microphone was utilized in the recording of data. Through this method, cost-effective and straightforward UAV payload detection can be conducted.

Audio data was recorded of two UAVs, DJI Phantom 4 and the Evo 2 pro, two dummy explosive payload were

utilized. The final collected data-set was 7.13 hours in total length, consisting of 2567 samples. Feature extraction and Data augmentation methods were conducted in the pre-processing stage. The results of pre-processing were put into the DL models; CNN, RNN, and CRNN. In the CNN, MFCC produced the best performance with an accuracy 94.93% . Mel presented an excellent result of 92.69% accuracy in CRNN. And MFCC has adequate accuracy of 81.13% with RNN. The limitation of this study is that only two UAVs were used to collect audio samples. Additionally, many more shapes and weights of payload could be analyzed. Future studies are expected to collect different kinds of data with manifold UAVs and payload to generalize the UAV payload detection system.

ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2019-0-01816)

REFERENCES

[1] M. al Kibsi", "Houthi drone targets senior Yemeni officers, kills five soldiers," Available at https://www.aljazeera.com/news/2019/1/10/houthi-drone-targets-senior-yemeni-officers-kills-five-soldiers (2022/05/18).

[2] W. Ripley", "Drone with radioactive material found on japanese prime minister's roof," Available at https://www.cnn.com/2015/04/22/asia/japan-prime-minister-rooftop-drone/index.html (2022/05/18).

[3] E. Kelly, "Venezuela drone attack: Here's what happened with nicolas maduro," *USA Today*. [Online]. Available: https://www.usatoday.com/story/news/politics/2018/08/06/venezuela-drone-attack-nicolas-maduro-assassination-attempt-what-happened/913096002

[4] Y. Wang, F. E. Fagiani, K. E. Ho, and E. T. Matson, "A feature engineering focused system for acoustic uav payload detection." in *ICAART (3)*, 2022, pp. 470–475.

[5] S. Wei, S. Zou, F. Liao *et al.*, "A comparison on data augmentation methods based on deep learning for audio classification," *2019 2nd International Conference on Computer Information Science and Artificial Intelligence(CISAI 2019)*, pp. 1–8, 2020.

[6] S. Al-Emadi, A. Al-Ali, A. Mohammad, and A. Al-Ali, "Audio based drone detection and identification using deep learning," *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, pp. 459–464, 2019.

[7] S. Jamil, M. Rahman, A. Ullah, S. Badnava, M. Forsat, S. S. Mirjavadi *et al.*, "Malicious uav detection using integrated audio and visual features for public safety applications," *Sensors*, vol. 20, no. 14, p. 3923, 2020.

[8] Y. Wang, F. E. Fagian, K. E. Ho, and E. T. Matson, "A feature engineering focused system for acoustic uav detection," *2021 Fifth IEEE International Conference on Robotic Computing (IRC)*, pp. 125–130, 2021.

[9] L. Shi, I. Ahmad, Y. He, and K. Chang, "Hidden markov model based drone sound recognition using mfcc technique in practical noisy environments," *Journal of Communications and Networks*, vol. 20, no. 5, pp. 509–518, 2018.

[10] V. Kartashov, V. Oleynikov, I. Koryttsev, S. Sheiko, O. Zubkov, S. Babkin, and I. Selieznov, "Use of acoustic signature for detection, recognition and direction finding of small unmanned aerial vehicles," *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, pp. 1–4, 2020.

[11] U. Seidaliyeva, M. Alduraibi, L. Ilipbayeva, and A. Almagambetov, "Detection of loaded and unloaded uav using deep neural network," *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pp. 490–494, 2020.

[12] U. Seidaliyeva, M. Alduraibi, L. Ilipbayeva, and N. Smailov, "Deep residual neural network-based classification of loaded and unloaded uav images," *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pp. 465–469, 2020.

[13] F. Fioranelli, M. Ritchie, H. Griffiths, and H. Borrion, "Classification of loaded/unloaded micro-drones using multistatic radar," *Electronics Letters*, vol. 51, no. 22, pp. 1813–1815, 2015.

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

[15] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[16] V. Tiwari, "Mfcc and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[17] K. M. Rashid and J. Louis, "Times-series data augmentation and deep learning for construction equipment activity recognition," *Advanced Engineering Informatics*, p. 100944, 2019.

[18] M. Esa, N. Mustaffa, H. Omar, N. M. Radzi, and R. Sallehuddin, "Learning convolution neural network with shift pitching based data augmentation for vibration analysis," vol. 864, no. 1, p. 012086, 2020.

[19] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[20] 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," pp. 131–135, 2017.

[21] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," pp. 1310–1318, 2013.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] S. H. Bae, I. K. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of lstm and cnn." pp. 11–15, 2016.

[24] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, "Heart sound classification based on improved mfcc features and convolutional recurrent neural networks," *Neural Networks*, vol. 130, pp. 22–32, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608020302306