Deep Learning Models for Gunshot Detection in the Albufera Natural Park

N. P. García-de-la-Puente[†], F. Fuentes-Hurtado[⋆‡], L. Fuster[⋄], V. Naranjo[†],G. Piñero[⋄]
[†] I3B, Universitat Politècnica de València, Spain, vnaranjo@dcom.upv.es
[⋆] Dep. Sistemas Informáticos, Universidad Politécnica de Madrid, Spain
[‡] KNODIS Research Group, Universidad Politécnica de Madrid, Spain, felix.fuentes@upm.es
[⋄] ITEAM, Universitat Politècnica de València, Spain, gpinyero@iteam.upv.es

***Abstract*—Gunshot detection in natural environments is crucial for the protection of endangered species. In this work, we present a novel dataset built from the soundscape recording at five different locations of the Spanish Albufera National Park. We then carry out an experimental study to detect gunshots from the rest of the background sounds and noises labeled as "background". For this purpose, we perform comprehensive experiments both in the data input and the modeling stages of three efficient deep convolutional neural networks (DCNNs), obtaining an F1 score (harmonic mean of precision and recall) of 0.92 for the best model. The best three DCNNs are also used to monitor one hour of the Albufera soundscape where the gunshot class represents 8% of the testset. The recall values obtained with our model are comparable to previous works monitoring gunshots in real scenarios.**

***Index Terms*—Gunshot detection, natural environment, deep learning, convolutional neural networks.**

I. Introduction

Gunshot detection in natural environments is of great importance for the protection of endangered species. Automatic systems capable to carry out gunshot detection have been researched for many years. The first systems were developed within the framework of multimedia sensor networks [1], focusing on the deployment of the recording devices and the communication among them, while in recent years research has focused on algorithms to detect gunshots in real environments as nature and cities [2]–[4].

Our work consists in a preliminary study of different deep neural network models to detect gunshots in the environment of the Albufera Natural Park (Spain), where an installation of audio devices described in Section II-B cover the lagoon area of the park. Our goal is to find out which deep learning models can efficiently model gunshots in real environmental conditions.

In recent decades, sound event detection (SED) solutions have been proposed to predict the occurrence of certain events in audio signals [5], [6]. Their approach is similar to that of performing a sequence labeling task: at each step, it has to distinguish the presence or absence of a particular event. The main problems that a sound event detector has to overcome are the variability of acoustic environments, such as background noise and distance to the sound, the appropriate choice of the size of the temporal processing window, and the presence of other overlapping sounds.

With respect to previous work on gunshot detection using Deep Learning (DL) architectures, [7] uses the combination of several feature matrices in a single-color image. Their algorithm was tested with noisy audio recordings and proved to be robust against false positive detections. In [8], a new architecture based on standard residual blocks is proposed and evaluated using a real audio track from an action movie. Recently, attention mechanisms [9] have also been applied to this problem. A detailed discussion on previous works achievement in real life environment condition will be provided in Section IV-A.

Our significant contributions to the field of acoustic event classification are[1]:

- We build a preliminary dataset of the soundscape of the Albufera Natural Park formed by two classes of audio events: "gunshot" and "background".
- We identify appropriate parameters of the preprocessing stage and the DL models for gunshot detection in the park through extensive experimentation, providing valuable insight for researchers working with similar types of audio data.
- We obtain comparative results to previous works on gunshot detection using similar real environment datasets.

II. Audio dataset

A. Study area

The Albufera Natural Park is located on the Gulf of Valencia coast in eastern Spain ($39°17'$ N, $00°20'$ E), and has a surface area of $20,956$ ha. The park has a large shallow water lagoon extended over $23.94\,\mathrm{km}^2$ called "L'Albufera", which is fed by streams, rivers, and irrigation channels, and is one of the most representative and valuable coastal wetlands of the Valencian Community and the Mediterranean basin. It was declared a Natural Park by the Spanish Government in 1986, and since 1989 it has been recognized as a "Wetland of International Importance"by the "The Convention on Wetlands"[2]. It is also

This work has been partially funded by Junta de Andalucía through Grant P20_01078, by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" through Grant PID2021-124280OB-C21, and by EU Horizon Europe through Grant Agreement No 101057404.

[1]Python code underlying DL models and gunshot dataset is available on GitHub https://github.com/gpinyero/EUSIPCO2023.

[2]https://rsis.ramsar.org/ris/454

Fig. 1. Map of the L'Albufera lagoon and surrounds with the location of the acoustic nodes.

an integral part of the European Natura 2000 Network, having been declared a "Special Protection Area for Bird" (SPA) in 1990. The overall waterbird numbers reach a mean of $80,000$ annual individuals.

Even though there are 45 endangered animal species in the area, many of them waterbirds, hunting is permitted in restricted areas inside the natural park. This year's hunting season began on November 13, 2022, and ended on February 12, 2023, with hunting allowed only on Saturdays, Sundays, and holidays, except during the third week of January, when hunting was also allowed on weekdays. In this sense, automatic gunshot detection is very important for the technical staff of the park for:

- Detecting poachers on prohibited days.
- Localizing hunting in the proximity of the Albufera bird nature reserves.

B. Recording techniques

Ten acoustic nodes have been deployed in the lagoon area, as shown in Fig. 1. The acoustic nodes are commercial devices, specifically the "Song Meter Mini" model from *Wildlife Acoustics*, whose specifications are available on their website.[3]. Five nodes were installed in four small islands inside the lagoon (*Mata de l'Antina*, *Mata de San Roc* (2 nodes), *La Maseguerota* and *Mata del Fang*) and the other five nodes were installed on the land surrounding the lagoon: *Tancat de Milia* (2 nodes), *La Tancadeta* (2 nodes) and the park's administrative office. The three places can be accessed only by authorized staff. In all the places with two nodes, the devices were at least 80 m apart one from another. The sample rate of all the nodes was set to $f_s = 24000$ Hz.

The overall goal of the installation is not only to detect dangerous situations, but mainly to study the birds' behavior in the Albufera environment. For this reason, the acoustic devices were programmed to record the soundscape when the birds are more active, thus, before and after the sunrise and the sunset. However, hunters usually go out during the sunrise period, so we will use for this study the recordings taken from 5.30 a.m. to 10.30 a.m. CET. At this moment we have access to the soundscapes recorded between November 13 and December 14, 2022, saved as wav files of 60 minutes duration each, resulting in 160 hours of audio recordings.

[3]https://www.wildlifeacoustics.com/products/song-meter-mini

C. Gunshot dataset

Only the two nodes located at *Tancat de Milia*, two nodes at *La Tancadeta*, and the node from *La Maseguerota* island have been used to build the gunshot dataset used in this work. The land nodes are located close to grounds where hunting is permitted, so a reasonable amount of gunshots could be found, especially from 6.30 a.m. in advance.

The background sounds that can be found in the recordings contains a wide set of natural sounds (birds and frogs singing, noises due to animals moving around the node, wind noise) together with human-caused noises such as airplanes, boat engines, speech, etc. On the East and West sides of the park, there are two roads with intense traffic during weekdays, but the five selected nodes are far enough to get affected.

As said before, we collected 160 hours of audio recording. However, to build the dataset, we analyzed only the first ten minutes of the five nodes recorded on November 13 at 8:30 a.m. Therefore, the training dataset consists of 418 audio segments that include one or more gunshots, labeled as "gunshot", and 449 audio segments that do not include any, labeled as "background" [10]. All of them have been extracted from recordings taken at the two nodes of *Tancat de Milia* and the two nodes of *La Tancadeta* and have a duration of 3 seconds. From the whole training set of 867 segments, 15% of the segments are used for validation.

The test dataset has been extracted from the *La Maseguerota*'s node, including 55 "gunshot" and 171 "background" audio segments of 3 seconds each. We have listened to other audio recordings taken on different days and hours, and we can conclude that the built dataset is representative of the two classes. More information about the database content will be discussed in Section IV-A.

III. Method

A. Data processing

The Mel spectrogram has been widely used when dealing with the problem of sound event detection in natural environments [10], [11] and it is also used here. We compute the 128-bands Mel spectrogram of 3 seconds segments using frames of 900 samples (37.5 ms), weighted by a Hanning window of the same length and with a 50% frame overlap, covering the frequency range $0 - 12.000$ Hz. As a consequence, each 3-second segment has been converted to a time-frequency matrix of 159×128 real values. As a final step, the decimal logarithm of the matrix elements has been computed.

Moreover, to find out the relevance of the frequency range in gunshot detection, we carried out different experiments varying the number of bands. We evaluate our models with the full spectrogram of 128 bands, but also with a cropped

spectrogram containing the first 83 bands ($0 - 3.818$ Hz) and only the first 60 bands ($0 - 1.995$ Hz).

Finally, since the DL models employed in this study use a ReLU activation function, it would be convenient to normalize the input elements between 0 and 1 since the ReLU activation function zeroes out the negative values. For this purpose, we computed the log-magnitude with two different reference levels in the *librosa.power_to_db* python function: 1) the maximum absolute value of the Mel spectrogram, and 2) the mean of their median values. After computing the log-magnitude, we applied an optional normalization between 0 and 1.

B. Models

We employed three different Deep Convolutional Neural Networks (DCNN) architectures in our experiments: VGG-16 [12], ResNet-34 [13] and MobileNet v2 [14].

1) VGG: VGG16 is a DCNN that consists of 16 layers, including 13 convolutional layers and 3 fully connected layers. The convolutional layers are arranged in sequential blocks, with each block consisting of multiple convolutional layers followed by a max-pooling layer. These convolutional layers extract features from the input audio signal at increasing levels of abstraction. The first layer learns low-level features, such as edges and curves, while subsequent layers learn more complex features, such as shapes and patterns. The final layers of the VGG16 architecture consist of fully connected layers, which combine the extracted features to make the final classification decision.

2) MobileNet v2: The MobileNet v2 architecture has demonstrated strong performance in image classification tasks while requiring fewer computational resources compared to other architectures. MobileNet V2 is a lightweight CNN.

The MobileNet v2 architecture consists of several building blocks, each containing a depthwise convolutional layer followed by a pointwise convolutional layer. The depthwise convolutional layer applies a separate convolutional filter to each channel of the input audio signal, while the pointwise convolutional layer applies a 1x1 convolutional filter to combine the features from the depthwise convolutional layer. This separation of the convolutional layers reduces the computational cost of the model and allows for faster training and inference.

3) ResNet-34: ResNet-34 is a DCNN that has been shown to perform well on a wide range of image classification tasks. The architecture consists of 50 layers, with skip connections that allow for deeper networks without suffering from the vanishing gradient problem.

The input to the ResNet-34 model is passed through a series of convolutional layers with varying kernel sizes and numbers of filters. The output of each convolutional layer is passed through a batch normalization layer and a rectified linear unit (ReLU) activation function. The network also includes four "residual blocks" that use skip connections to improve gradient flow and enable deeper networks. The final output of the ResNet-34 model is a vector of probabilities indicating the likelihood that the input audio clip contains a gunshot.

In our implementation, we modified its architecture by replacing the last fully connected layer with a sigmoid layer to obtain a binary classification output.

C. Activation functions

LeakyReLU (Leaky Rectified Linear Unit) is an activation function used in neural networks, which is similar to the ReLU (Rectified Linear Unit) activation function but allows a small, non-zero gradient when the input is negative.

The ReLU function is defined as

$$f(x) = \max(0, x)\,,$$

with 0 output when the input x is negative and x otherwise. While ReLU has been widely used in deep learning models due to its simplicity and computational efficiency, it can suffer from the "dying ReLU" problem where the gradient becomes zero for negative inputs and the neuron effectively becomes inactive. To address this issue, the LeakyReLU was introduced, which modifies the ReLU function to

$$f(x) = \max(ax, x)\,,$$

where a is a small, positive constant (typically 0.01).

This means that for negative inputs, the function outputs a small, non-zero value instead of 0, which allows the gradient to flow even for negative inputs. Regarding our problem, the Mel spectrogram can present negative and positive values after computing the logarithm, thus it is interesting to consider the LeakyReLU activation function. Additionally, it has been shown to be more robust to noisy input data and can help to prevent overfitting [15].

D. Loss function

The binary cross-entropy loss function is commonly used in binary classification tasks, such as gunshot detection. It measures the difference between the predicted probability distribution and the true probability distribution of the target class.

The binary cross-entropy loss function is defined as:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})\,, \quad (1)$$

where y is the true binary label (0 or 1), $\hat{y}$ is the predicted probability of the positive class (i.e., the probability of a gunshot sound in our case), N is the total number of examples, and $\log$ is the natural logarithm. The loss function penalizes the model when it predicts a low probability for the true positive class and when it predicts a high probability for the true negative class.

E. Evaluation Metrics

The output of our model gives the probability of detecting a gunshot in a given audio segment. We use a threshold of 0.5 to assume that current 3-seconds audio segment contains one or more gunshots ($p > 0.5$). Using these predictions, we compute several metrics to assess the performance of the classifier, namely: precision, recall, and f1 score:

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

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

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

For detecting rare sound events, the goal is to obtain a balanced trade-off between recall (the proportion of true positive events that are correctly identified) and precision (the proportion of identified positive events that are true positives). However, the F1 score is a significant metric when the rare event is critically important to detect, as it takes into account both precision and recall values and provides a balanced measure of the model's performance [16].

F. Experimental setup

We trained all our models using the dataset described in section II during 85 epochs with the Adam optimizer and a learning rate of 0.0001. An adaptive learning rate scheduler called *Reduce LR on Plateau* is used. Thanks to the patience parameter (set to 10 in our experiment), it determines how many epochs the algorithm waits before reducing the learning rate. The categorical cross-entropy loss was computed for a batch size of 32.

Regarding the hardware, we used NVIDIA RTX 3090 24 GB x 1, 525.60.11 drivers & CUDA 12.0, MSI Z270 Gamming PRO Carbon (MS-7A63); 32 GB and Intel i7-7700K (4.2 GHz), whereas the software used was Pytorch and Torchaudio for building and training the models, and Sci-kit learn for evaluation.

IV. Experimental Results

Table I shows the set of conditions for the three models (VGG16, ResNet-34 and MobileNet V2) regarding the reference values and the activation functions as described in Sections III-A and III-C, respectively. In addition, we tested the three sizes of the Mel spectrograms regarding their frequency axis (128, 83, and 60) as explained in Section III-A. Therefore, we run $6 \times 3 \times 3$ training experiments, with a total of 54 models. We then selected the top-5 performing configurations for each DCNN, whose metrics obtained on the test dataset are shown in Table II.

It can be noted from Table II that VGG16 shows the best performance for the reduced frequency range of 60 and 83 bands, and considering the average median of their input values as a reference. However, ResNet34 presents good results with the full 128-bands image in three of their top five models, although its performance is poorer than VGG16.

TABLE I
SET OF CONDITIONS FOR THE THREE NNS

exp	reference	norm	leaky
1	max	False	False
2	max	False	True
3	max	True	False
4	max	True	True
5	medians	True	False
6	medians	True	True

TABLE II
TOP-5 CONFIGURATIONS FOR EACH MODEL (TEST SET).

model	exp	crop	*precision*	*recall*	*f1score*
$VGG16_1$	6	60	0.87	0.98	**0.92**
$VGG16_2$	5	83	0.88	0.96	**0.92**
$VGG16_3$	6	83	0.84	0.98	**0.91**
$VGG16_4$	4	128	0.86	0.93	0.89
$VGG16_5$	5	60	0.8	0.96	0.88

(a) VGG16 model

model	exp	crop	*precision*	*recall*	*f1score*
$MobileNetV2_1$	3	60	0.46	0.67	0.55
$MobileNetV2_2$	1	83	0.36	0.85	0.51
$MobileNetV2_3$	3	83	0.37	0.76	0.49
$MobileNetV2_4$	2	60	0.45	0.51	0.48
$MobileNetV2_5$	6	83	0.44	0.47	0.46

(b) MobileNet V2 model

model	exp	crop	*precision*	*recall*	*f1score*
$ResNet34_1$	1	128	0.85	0.93	0.89
$ResNet34_2$	5	83	0.86	0.87	0.86
$ResNet34_3$	2	60	0.8	0.89	0.84
$ResNet34_4$	5	128	0.85	0.84	0.84
$ResNet34_5$	2	128	0.8	0.89	0.84

(c) ResNet34 model

Regarding MobilenNetV2, it appears that it is not properly working, possibly because of their lower number of parameters with respect to the other two DCNNs.

Comparing to previous work on gunshot detection, our best results lie very close to the range of recall and precision found in the literature, even if our results are for the test dataset and most of the previous works obtain their metrics from the validation dataset. In [16], they used a ResNet18 CNN architecture and obtained a recall of 95% and a precision of 85% for a real dataset recorded in the Latin American forest. In [7], they used a synthetic dataset of gunshots mixed with background sounds such as traffic noise, human voice, animal sounds and other forms of environmental sounds. Its best model achieved a recall of 97.6% for the gunshot class with a Resnet18 CNN, while its precision was near to one. In [4], they used several 1-D and 2-D CNN models with a majority-rules ensemble, obtaining an accuracy above 99% on validation data for a residential environment in Santa Clara, CA.

A. On gunshot detection in real conditions

To assess the DCNNs' performance when working in real conditions, that is, monitoring the audio recordings to detect gunshots, we selected the three best performing models marked in bold in Table II to predict an audio excerpt from

TABLE III
RECALL FOR THE TOP-3 MODELS UNDER REAL MONITORING

model	$VGG16_1$	$VGG16_2$	$VGG16_3$
gunshot	0.42	0.52	0.49
background	0.89	0.84	0.89

the *La Maseguerota*'s node of one hour duration. The test was performed every second, extracting the segment containing the following 3 seconds and labeling it as "gunshot" if any gunshot is detected by the DL model, as it would be performed in real-life conditions. Therefore, the number of total events presented to the DCNNs were 3595, from which the groundtruth consist on 285 labeled as "gunshot" and 3310 as "background". In some cases, the gunshots were very far from the recording node and hard to distinguish by the human ear, to say that gunshots appear with very different levels of signal-to-noise ratio (SNR) with respect to the background soundscape. Since the testset is very unbalanced, we show the obtained recall values for the two classes in Table III.

Comparing to previous work, [16] obtained a 57% of correctly identified gunshots in the Latin America forest when monitoring new recordings. However, they added the false positives of the "gunshot" class obtained from the new recordings to the background training data, in order to help the CNN to learn background sounds that have similar properties to gunshots. This is a further step that we want to explore with new node's recordings as well. A poorer result of 35 identified gunshots from 280 was obtained by [4] when their model was tested in Indianapolis. After retraining the original model (obtained from a residential dataset) on this data, 158 out of 342 audio clips containing gunshots were positively identified, resulting in a recall of 46.2%.

Finally, Fig. 2 shows clockwise four samples of Mel spectrograms detected as true positive (TP), false positive (FP), true negative (TN) and false negative (FN) to notice how some background sounds as duck quacking can be easily confused with gunshots.

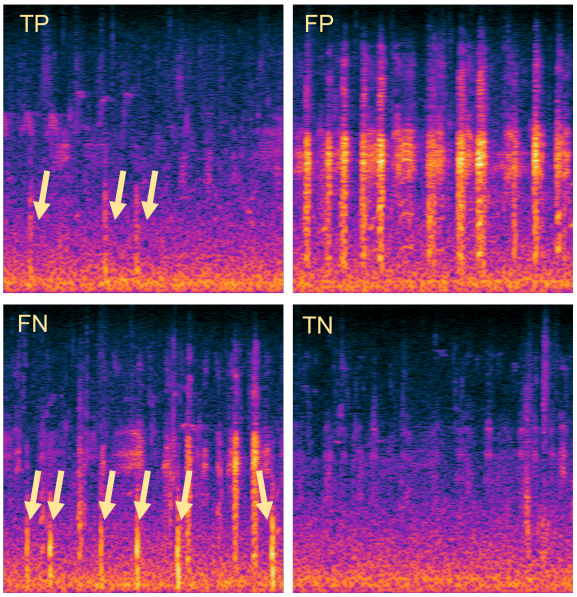

Fig. 2. True positive (top left, far gunshots are detected), false positive (top right, quacking of a duck), false negative (bottom left, not detected gunshots combined with a quacking) and true negative (bottom right).

V. CONCLUSIONS

We have carried out an experimental study on a novel dataset built from the soundscape recording at five different locations of the Spanish Albufera National Park. Our goal was to detect gunshots from the rest of the background sounds and noises labeled as "background". For this purpose, three deep CNNs have been trained for different input and model conditions resulting in 54 different networks. The best three CNNs have also being used to monitor the soundscape in real conditions, obtaining comparable results of recall values with respect to previous works evaluating real datasets of Latin American Forests and American cities. In the future, our aim is to propose new approaches that can improve the F1-score, such as anomaly detection techniques.

REFERENCES

[1] I.F. Akyildiz, T. Melodia, and K.R. Chowdhury, "A survey on wireless multimedia sensor networks," *Computer Networks*, vol. 51, pp. 921–960, 2007.

[2] A.P. Hill *et al.*, "Audiomoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment," *Methods in Ecology and Evolution*, vol. 9, pp. 1199–1211, 2018.

[3] J. Salamon, C. Jacoby, and J.P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM International Conf. on Multimedia - MM '14*, 2014, pp. 1041–1044.

[4] A. Morehead, L. Ogden, G. Magee, R. Hosler, B. White, and G. Mohler, "Low cost gunshot detection using deep learning on the raspberry pi," in *Proc. 2019 IEEE Int. Conf. on Big Data*, 2019, pp. 3038–3044.

[5] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia Computer Science*, vol. 112, pp. 2048–2056, 2017.

[6] A. Mesaros, T. Heittola, T. Virtanen, and M.D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, pp. 67–83, 2021.

[7] J. Bajzik, J. Prinosil, and D. Koniar, "Gunshot detection using convolutional neural networks," in *2020 24th Int. Conf. Electronics*, 2020, pp. 1–5.

[8] J. Bajzik, J. Prinosil, R. Jarina, and J. Mekyska, "Independent channel residual convolutional network for gunshot detection," *Int. Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, 2022.

[9] R. Nijhawan, S.A. Ansari, S. Kumar, F. Alassery, and S.M. El-Kenawy, "Gun identification from gunshot audios for secure public places using transformer learning," *Scientific Reports*, vol. 12, pp. 13300, 2022.

[10] D. Stowell, T. Petrusková, M. Šálek, and P. Linhart, "Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions," *J. R. Soc. Interface*, p. 20180940, 2019.

[11] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, pp. e13152, 3 2022.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[15] A.L. Maas, A.Y. Hannun, A.Y. Ng, et al., "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30.

[16] L.K.D. Katsis *et al.*, "Automated detection of gunshots in tropical forests using convolutional neural networks," *Ecological Indicators*, vol. 141, pp. 109128, 8 2022.