Sound-based logging detection using deep learning

Budimir Anđelić
University of Montenegro
Faculty of Electrical Engineering
Podgorica, Montenegro
budimir.andjelic@gmail.com

Milutin Radonjić
University of Montenegro
Faculty of Electrical Engineering
Podgorica, Montenegro
mico@ucg.ac.me

Slobodan Đukanović
University of Montenegro
Faculty of Electrical Engineering
Podgorica, Montenegro
slobdj@ucg.ac.me

***Abstract*— Illegal logging represents a major environmental issue which impedes the stability of forest ecosystem and supports climate change, flooding, soil erosion, weeding of habitats, extinction of animal and plant species. This paper proposes a method for mitigating the logging issue by automatically detecting the sound of logging activities. More specifically, we propose to detect the chainsaw sound using deep learning. Two deep learning approaches were considered, one based on multilayer perception (MLP) and the other based on convolutional neural network (CNN). As inputs to our models, we used time, frequency and time-frequency audio features. For the purpose of this research, we collected two datasets of audio signals. First dataset, downloaded from YouTube, is used for training and validating the proposed models. Second dataset, which we recorded in real conditions, is used for testing the proposed models. The experiments show that the CNN-based approach outperforms the MLP-based one, with a sound classification accuracy of 94.96% on the first dataset and 88.87% on the second dataset.**

Keywords—Logging, chainsaw, deep learning, audio features, mel-frequency cepstral coefficients, mel spectrogram

I. INTRODUCTION

Forests represent one of the most important natural resources, which, in addition to the primary task of wood production, provide protection from harmful gases, oxygen production, protection of human health, habitat for animals, soil protection, protection from wind and floods, tourism resources, etc. Although the survival of mankind largely depends on the health of forests, their destruction continues by day. Besides forest fires, serious threat to forest ecosystems is illegal logging, which disrupts the stability of the ecosystem, with consequences reflected in weeding of habitats, drying of trees, broken trees, soil erosion, climate change, disappearing of many plant and animal species, etc. In addition to the long-term consequences for the environment and people's quality of life, illegal logging also brings great economic losses.

According to [1], forests and forest land make up 69.4% of the territory of Montenegro, one of the highest percentage in Europe. Based on the reports of the Ministry of Agriculture and Rural Development, an average of 5,680 m^3 of wood is illegally cut every year in Montenegro, of which only 8.13% is confiscated. However, this problem is much more serious in reality, as evidenced by field visits. The fact that only the high-quality trees are cut, contributes to the negative impact on the condition of forests. There are 422 employees in the Forestry Administration, which is responsible for forest management in Montenegro, of which 70 are engineers, 171 forest guards and 56 forestry technicians [1]. According to these data, each engineer covers about 140 km^2, and each guard about 56 km^2. In addition to the lack of manpower, there is also a lack of adequate equipment and vehicles to prevent illegal logging.

This study proposes a system that will help in preventing illegal logging. It is organized as follows. Section 2 provides an overview of previous research on the topic of chainsaw sound detection. Section 3 discusses the methodology used for illegal logging detection based on audio features and neural networks. Section 4 describes the datasets used for neural network training and real-world testing. Section 5 presents the obtained results, and, finally, section 6 concludes the paper.

II. OVERVIEW OF RELATED WORKS

In [2], a system for automatic chainsaw sound detection was proposed in order to prevent illegal logging in the Amazon rainforest. The system was developed in the form of sensor nodes that are part of a wireless acoustic sensor network. Each sound is represented by a set of mel-frequency cepstral coefficients (MFCC), and a probability density function (PDF) is fitted with a kernel based on a multivariate Gaussian density estimation using only the target class. The method enables recognition of only chainsaw sounds rejecting other environmental sounds (animal's calls, weather noises or boat engines). The classification accuracy achieved is 98%. The dataset contains 560 samples related to the sound of a chainsaw, 5472 to environmental sounds, and 5363 samples related to the sounds of engines and human speech.

As a solution to the problem of illegal tree cutting, the authors in [3] proposed a technique based on the extraction of 2D Haar wavelet coefficients from the sound spectrogram. Two different approaches were followed. The first approach applies two decision thresholds to the extracted features, whereas the second one performs binary classification using the support vector machine (SVM). The achieved accuracy reaches 97%. However, there are specific sounds that have a large classification error. Sounds produced by animals such as cows and deer, some insects and human voice have a spectrum with a high degree of similarity to a chainsaw. Chainsaw sounds make up 35% of the dataset, while sounds similar to chainsaw make up 3% of the dataset.

Paper [4] proposes an alternative scheme for extraction of chainsaw sound features based on MFCC and sinusoidal lifter. Although being a proper choice when it comes to classifying audio signals, MFCC underperforms in the presence of noise. In [4], the Mel coefficients are passed through a sinusoidal lifter to solve the noise problem. The experiments show that the lifter increases the learning capacity of a neural network compared to the MFCC without the lifter. The ESC-50 dataset [5] was used. It consists of 50 classes, and each class contains 40 5-sec audio recordings. The achieved accuracy using MFCC without a lifter is 36.37% and with a lifter 56.67%.

Study [6] focused on the possibility of detecting illegal logging based on the chainsaw recognition by combining MFCC with k-nearest neighbor (*k*-NN) and SVM. The dataset comprises 3265 5-sec audio signals, with less than 10% related to the chainsaw sounds. An accuracy of 95.63% was achieved with SVM and 94.02% with *k*-NN. Although the accuracies are approximate, the SVM algorithm proved to be a much better option due to shorter processing time. It required 30 times less time to process the sound. However, the accuracy drops drastically (53.16% for SVM, 40.50% for *k*-NN) if there are sounds similar to the chainsaw.

In [7], an algorithm for detecting the chainsaw sound based on a sensor node with limited energy is presented. It uses three techniques: adaptive energy threshold, delta pitch detection and energy band ratio. The dataset consisted of 250 audio signals. The achieved accuracy is 90.8%.

Related to the data used for training the models for illegal logging detection, it is important to point out that only few of them use an adequate dataset for training and testing. The classes are not properly balanced, so the training methods consider one class more important than others. In addition, the authors often do not specify which data were used for testing, as well as whether the training and testing data are from different sources, which puts the achieved accuracies in question.

III. Fundamentals

Machine learning represents an effective framework to prevent illegal logging by detecting the sound of chainsaw [3]-[7], the most often used tool for cutting trees. It is very important to take into account the environmental sounds (rain, thunderstorm, wind, walking, crickets, bird's chirping, owl howling), as well as the sounds similar to chainsaw (grass trimmer, lawn mower, dirt bike, snowmobile) in order to avoid false alarms. We further present several standard audio features used in machine learning-based sound classification.

A. Audio features

Audio features can be divided into three domain-based categories: time-domain, frequency-domain and time-frequency domain audio features. Time domain audio features are extracted from the sound waveform. In this work, the following time domain features were used: amplitude envelope (AE), root-mean square energy (RMSE) and zero crossing rate (ZCR). All these features provide instantaneous information about the sound.

AE describes how the sound changes over time [8] and represents the maximum amplitude of all samples in a frame:

$$\mathrm{AE}_t = \max_{tK \le k \le (t+1)K-1} s(k)\,, \tag{1}$$

where t is the frame number and K is the frame size. AE gives a rough information about the sound loudness, but it is sensitive to outliers. It is mainly used for onset detection or music genre classification.

RMSE is calculated as [8]:

$$\mathrm{RMSE}_t = \sqrt{\frac{1}{K}\sum_{k=tK}^{(t+1)K-1} s(k)^2}. \tag{2}$$

Like AE, RMSE is also an indicator of loudness and is often used for genre recognition and audio classification. Because RMSE is calculated based on the energy values of all samples in the frame, it is less sensitive to outliers.

ZCR is the number of times the signal crosses the horizontal axis [8]. It is defined as:

$$\mathrm{ZCR}_t = \frac{1}{2}\sum_{k=t\cdot K}^{(t+1)K-1} |sgn(s(k)) - sgn(s(k+1))|, \tag{3}$$

where $sgn(\cdot)$ is the sign function. It is often used for recognition of percussive, pitch estimation, speech detection, etc.

Frequency domain audio features are extracted from the short-time Fourier transform (STFT) of the signal. Frequency domain audio features used in this study are spectral centroid (SC) and spectral bandwidth (SB).

SC is defined as the frequency range where the largest amount of energy is concentrated [8]. It is calculated as:

$$\mathrm{SC}_t = \frac{\sum_{n=1}^{N} m_t(n)\, n}{\sum_{n=1}^{N} m_t(n)}, \tag{4}$$

where n is the frequency bin, t is the number of the frame, $m_t(n)$ is the magnitude of the signal at frequency bin n and frame t, and N is the total number of frequency bins. This feature is associated with the measure of sound brightness (the presence of high frequencies) and is often used in digital audio signal processing to determine the sound color.

SB is derived from SC, and it refers to the spectral range of interesting parts in the signal [8]. It is calculated as the mean value of the distance between frequency ranges and SC:

$$\mathrm{SB}_t = \frac{\sum_{n=1}^{N} |n - SC_t|\, m_t(n)}{\sum_{n=1}^{N} m_t(n)}. \tag{5}$$

Time-frequency domain audio features combine both the time and frequency components of sound. The mel spectrogram and MFCC were used in this paper.

The mel spectrogram is obtained by applying the mel filter bank to the spectrogram of a signal. The number of mel bands is a parameter that can vary, depending on the problem.

MFCC are calculated through a discrete cosine transform of the mel spectrogram. MFCC is one of the most commonly used features when it comes to audio signal analysis, which provides information about the texture/color of the sound. Fig. 1 shows the mel spectrogram and MFCC of the chainsaw, bird's chirp and grass trimmer sound.

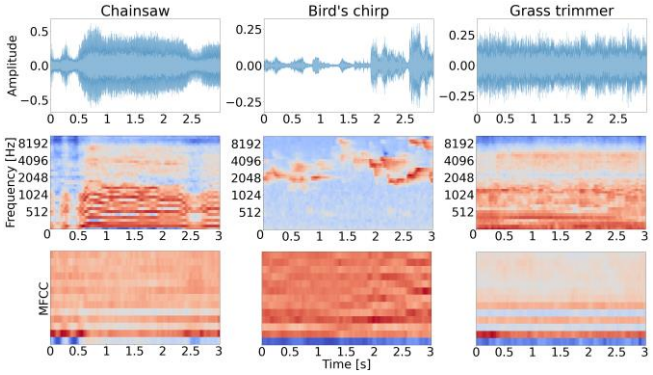

Fig. 1. Raw waveform (top row), mel spectrogram (middle row) and MFCC (bottom row) of the sounds of a chainsaw, bird's chirp and a grass trimmer.

B. Neural networks

In this study, two types of neural networks are used for chainsaw sound detection: multi-layer perception (MLP) neural network and convolutional neural network (CNN).

The MLP neural network architecture proposed in this study consists of an input layer, three hidden layers, and output layer. The number of neurons per hidden layers are 512, 256 and 64. In order to prevent overfitting, dropout, λ_2 regularization and early stopping were used. ReLU function was used as the activation function. The output layer contains

three neurons, corresponding to three categories of audio signals: chainsaw, similar to chainsaw and environmental. Softmax is the activation function of the output layer. The Adam algorithm is used for neural network weights optimization. It is recommended as default optimization algorithm because it is simple to implement, has faster training time, low memory requirements and requires less settings than any other optimization algorithm.

Sparse categorical cross-entropy (SCC) was used as the error function, which calculates the cross-entropy between actual outputs and predictions. Unlike the categorical cross-entropy, where the outputs are numbered binary, with SCC the outputs are integers (0, 1 or 2). In training, early stopping was implemented, which halts the training if the validation error does not decrease after 5 epochs.

The proposed CNN architecture consists of three convolutional layers, three pooling layers, one fully connected layer and one output layer. Type of convolutional layers is Conv2D. The number of kernels in convolutional layers is 32, the kernel dimensions are 3×3, the vertical and horizontal strides are 1 and the activation function is ReLU. Type of pooling layers is MaxPooling2D, pooling window size 3×3, and the vertical and horizontal stride is 2. In order to prevent overfitting, dropout and early stopping were used. In the last two fully-connected layers, the number of neurons is 64 and 3. The activation function in the output layer is Softmax.

For extracting audio signal features and implementing neural networks Python 3.7 programming language and PyCharm Community Edition 2021.3.2 IDE were used.

IV. Dataset

In this study, two datasets were collected. The first dataset, labeled as *Dataset* 1, consists of audio signals downloaded from the public platform www.youtube.com. Dataset 1 was used to train the neural networks and select the optimal model. In order to test the developed model in real conditions, another set of data, labeled as *Dataset* 2, was collected. It consists of audio signals that we recorded for the purpose of this research. Both datasets are available upon request. The programs used to prepare audio signals are Audacity [9] and WavePad Sound Editor [10].

Dataset 1 consists of 30-sec audio in wave audio format (*.wav), divided into training, validation and testing parts. During the preparation of the dataset, each audio signal is divided into 10 segments with a duration of 3 seconds. It is important to note that the audio signals used for training, validation and testing come from different sources.

Training, testing and validation data are then divided into three categories. The first category refers to environmental sounds such as: wind, rain, thunder, snow, river, lake waves, birds, other animals that are present in forests, walking in the forest, insects, etc. The second category refers to the sound of a chainsaw, of various manufacturers, at different stages of operation (starting the chainsaw, cutting the wood, idle), as well as at different distances. Also, this category is expanded with audio signals that represent a combination of chainsaw sounds and environmental sounds (e.g., chainsaw and thunder, chainsaw and rain, chainsaw and wind, etc.). The third category, *false positives*, refers to sounds that are very similar to a chainsaw: snowmobiles, motorcycles (two-stroke and four-stroke), grass trimmer, and lawn mower. This category is also expanded with audio signals that represent a combination of sounds very similar to a chainsaw and environmental sounds. All three categories are approximately equal in size.

The total number of audio signals, with duration of 30 seconds, in Dataset 1 is 9897, which means 98970 input signals after division into segments. 59520 inputs (60.14%) are used for neural network training, 19260 inputs (19.46%) are used for validation and 20190 inputs (20.40%) are used for testing. The variety of audio signals in Dataset 1, as well as the size itself (48.1GB or almost 100,000 input signals) will reduce the chances of neural network overfitting. The distribution of inputs in Dataset 1 is shown in TABLE I.

TABLE I. The distribution of inputs in Dataset 1

Category		Number of inputs		
		Train	***Test***	***Validation***
Chainsaw		19490	6560	6460
Environmental	Birds	4600	1340	1380
	Walking	760	270	270
	Wind	2850	770	770
	Rain and thunder	2990	1210	1260
	Insects	1410	600	590
	Night in forest	2560	760	710
	River and lake	2330	990	520
	Total	17500	5940	5500
False positive	Grass trimer	4740	1670	1590
	Lawn mower	6200	2050	1950
	Snowmobile	4970	1700	1590
	Motorcycles	6620	2270	2170
	Total	22530	7690	7300

Dataset 2 consists of audio signals with a duration of 30 seconds in wave audio format (*.wav). Audio signals are recorded under real conditions. Dataset 2 is divided into three categories: environmental sounds, chainsaw sounds and false positive. The total number of audio signals is 425, which means 4250 input signals after dividing into segments. The distribution of inputs in Dataset 2 is shown in TABLE II.

TABLE II. The distribution of inputs in Dataset 2

Category		Number of inputs
Chainsaw		1240
Environmental	Walking	165
	Wind	615
	Rain and thunder	565
	Insects	365
	Total	1710
False positive	Grass trimer	700
	Lawn mower	600
	Total	1300

V. Results

Inputs of the MLP neural network are time and frequency features: AE, ZCR, RMS, SC and SB. The features are concatenated one after another, and normalized by scaling between -1 and 1. This way, all features will be equally considered when training the neural network. The parameters that change when training the MLP neural network are: learning rate, dropout, λ_2 regularization parameter and batch

size. The learning rate takes values between 1e-5 and 1e-2. Dropout takes values 0.1, 0.2 and 0.3. The λ_2 regularization parameter takes values from 1e-7 to 1e-2. Batch size takes values from the set of elements: [8, 16, 32, 64, 128, 256, 512]. The number of epochs depends on the moment at which the early stopping is performed.

The best result of the MLP model on the test data of Dataset 1 was achieved for learning rate = 1e-4, batch size = 512, dropout = 0.2 and λ_2 = 1e-7. Out of total 20190 test inputs, 14299 of them or about 71% were correctly classified. The early stopping was performed after 22 epochs and the training time was 23.11 seconds. Fig. 2a) shows the confusion matrix of MLP neural network.

The inputs to CNN are time-frequency features: MFCC and mel spectrogram. The number of MFCC tested in this study is 13, 26 and 39. The number of mel bands in mel spectrogram tested is 40, 80 and 130. The variable parameters in training CNN are learning rate, dropout and batch size. The learning rate takes values between 1e-5 and 1e-2. Dropout takes values 0.1, 0.2 and 0.3. Batch size takes values from the set of elements: [64, 128, 256, 512]. The number of epochs is dictated by the implemented early stopping.

The parameters of the CNN model with the best result when MFCC are fed to the input are: learning rate = 1e-4, batch size = 256, dropout = 0.1 and the number of MFCC is 39. On the test data from Dataset 1, an accuracy of 92.39% was achieved. The early stopping was done after 14 epochs, and the training time was 1107.57 second. Fig. 2b) shows the confusion matrix of CNN for MFCC inputs.

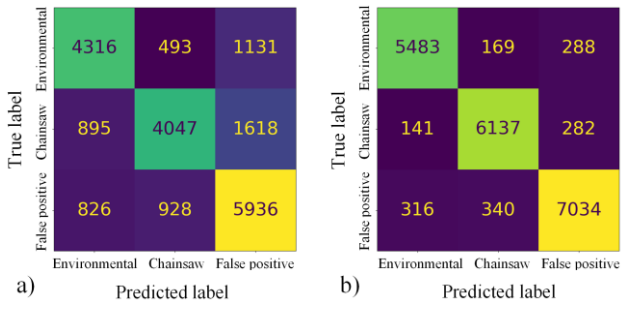

Fig. 2. Confusion matrix of a) MLP neural network b) CNN with MFCC inputs.

The parameters of the CNN model with the best result when the mel spectrogram is used as input are: learning rate = 1e-4, batch size = 512, dropout = 0.3 and the number of mel bands = 80. On the test data from Dataset 1, an accuracy of 94.96% was achieved. The early stopping was performed after 20 epochs and the training time was 3252 seconds. This is also the model that achieved the best result on the test data from Dataset 1 and will be selected as the optimal model. Fig. 3a) shows the confusion matrix for this CNN.

High accuracy on the test data from Dataset 1 shows that it is possible to detect the sound of a chainsaw, even in the presence of sounds that are very similar to a chainsaw. However, it is necessary to test the optimal model in real conditions on recorded audio signals. The optimal model trained on Dataset 1 achieved an accuracy of 88.87% on the recorded audio signals from Dataset 2. Fig. 3b) shows the confusion matrix of CNN on Dataset 2.

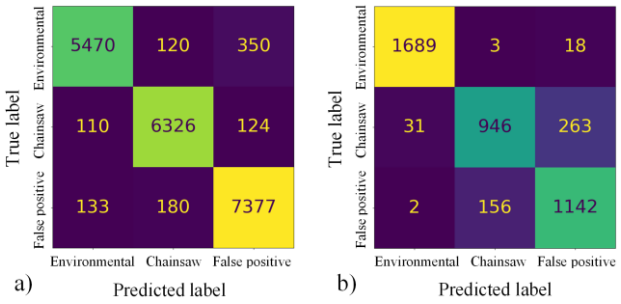

Fig. 3. Confusion matrix of a) CNN with mel spectrogram inputs b) CNN with mel spectrogram inputs tested on Dataset 2.

VI. Conclusion

In this paper, logging detection based on deep learning was addressed. Two models for chainsaw sound detection were proposed: MLP-based model and CNN-based model. The latter model, which proved to be more accurate, achieves the classification accuracy of 94.96% on a dataset collected from YouTube and 88.87% on a dataset recorded for the purpose of this research.

Future research will consider implementation of the proposed logging detection model on an affordable microprocessor platform equipped with a simple microphone, GPS and GSM/GPRS module. This system would be mounted in the forest and powered by a solar panel. When chainsaw detection occurs, the system will inform the user about event, with its exact coordinates. Also, such system with the appropriate dataset could find application in various fields.

References

[1] M. Gazdić, M. Gajević, D. Zarubica, and V. Iković. (2020). Šumarstvo, alternativa razvoja Crne Gore.; Available at: https://www.researchgate.net/publication/353795133_Sumarstvo_alternativa_razvoja_Crne_Gore.

[2] J. G. Colonna, B. Gatto, E. Miranda Dos Santos and E. F. Nakamura, "A Framework for Chainsaw Detection Using One-Class Kernel and Wireless Acoustic Sensor Networks into the Amazon Rainforest," 2016 17th IEEE International Conference on Mobile Data Management (MDM), 2016, pp. 34-36.

[3] G. Nicolae, A. Gaiță, A. Rădoi and C. Burileanu, "A Method for Chainsaw Sound Detection Based on Haar-like Features," 2018 41st International Conference on Telecommunications and Signal Processing (TSP), 2018, pp. 1-5.

[4] B. R. Ramadhan, M. Abdurohman and S. Prabowo, "Accuracy Enhancement of Feature Extraction Scheme in Detection of Chainsaw Sound to Prevent Illegal Logging," 2019 IEEE International Conference on Signals and Systems (ICSigSys), 2019, pp. 56-61.

[5] ESC-50 dataset; Available at: https://github.com/karolpiczak/ESC-50

[6] N. A. J. Gnamele et al., "KNN and SVM Classification for Chainsaw Sound Identification in the Forest Areas" International Journal of Advanced Computer Science and Applications (IJACSA), 10(12), 2019.

[7] T. Soisiinthorn and S. Rujipattanapong, "Deforestation detection algorithm for wireless sensor networks," 2007 International Symposium on Communications and Information Technologies, 2007, pp. 1413-1416.

[8] M. Müller. (2015). Fundamentals of Music Processing. 10.1007/978-3-319-21945-5.

[9] Audacity, https://www.audacityteam.org/download/

[10] Audio Editing Software. Sound, Music, Voice & MP3 Editor. Best Audio Editor for 2022 https://www.nch.com.au/wavepad/index.html