Independent Channel Residual Convolutional Network for Gunshot Detection

Jakub Bajzik[1], Jiri Prinosil[2], Roman Jarina[3], Jiri Mekyska[4]
Dept. of Mechatronics and Electronics, University of Zilina
Zilina 010 26, Slovakia[1]
Dept. of Telecommunications, Brno University of Technology
601 90 Brno, Czech Republic[2,4]
Dept. of Multimedia and Information and Communication Technology
University of Zilina, Zilina 010 26, Slovakia[3]

***Abstract*—The main purpose of this work is to propose a robust approach for dangerous sound events detection (e.g. gunshots) to improve recent surveillance systems. Despite the fact that the detection and classification of different sound events has a long history in signal processing, the analysis of environmental sounds is still challenging. The most recent works aim to prefer the time-frequency 2-D representation of sound as input to feed convolutional neural networks. This paper includes an analysis of known architectures as well as a newly proposed Independent Channel Residual Convolutional Network architecture based on standard residual blocks. Our approach consists of processing three different types of features in the individual channels. The UrbanSound8k and the Free Firearm Sound Library audio datasets are used for training and testing data generation, achieving a 98 % F1 score. The model was also evaluated in the wild using manually annotated movie audio track, achieving a 44 % F1 score, which is not too high but still better than other state-of-the-art techniques.**

***Keywords*—Acoustic signal processing; gunshot detection systems; audio signal analysis; machine learning; deep learning; residual networks**

I. Introduction

In the field of signal processing, the audio data analysis takes an extensive part, which is constantly studied. Many machine learning-based algorithms were proposed for solving tasks such as classification, segmentation, and denoising. In many cases, the methods are adapted for a specific type of sound, mainly speech and music, which take an extensive part in the research. On the other hand, the environmental sounds are unstructured, and it is challenging to generalize their nature. Many environmental sound analysis applications, ranging from urban monitoring [1] to IoT [2] and surveillance system [3], [4], [5], have been developed within the past years.

However, in recent years it has become a topical task to use known techniques to classify environmental sounds like explosion, gunshot, siren, car alarm, baby crying, window breakage and other events associated with potential danger [6]. Usage of learning algorithms tends to increase personal safety. The possible implementations are in-home or industrial protection systems, in cars to alert deaf or poorly hearing drivers of the siren, in homes to alert a parent of a crying child, and in a wide range of assistive devices, especially for deaf people. Recent modern surveillance systems for risk prevention purposes focuses mainly on the analysis of video signals from cameras using advanced computer vision techniques [7], [8], [9]. However, the analysis of audio signals has considerable potential in these systems as well. Especially gunshot detection technologies have been increasingly adopted by law enforcement agencies for mapping the spatial and temporal patterns of gun violence [10]. The general problem of gun detection technologies is a high rate of false alarms resulting in the waste of police resources when responding to those false alerts [11]. From a practical point of view, it is necessary to minimize the amount of false-positive predictions. Therefore, when setting the operating point of the system in practical applications, the false-positive rate must be taken into account.

The main objective of our work is to propose a method for detecting danger-related audio events (gunshots) that achieve high specificity in real conditions. We also aim to explore several types of feature spaces and neural architectures and discuss, what kind of setup is the most suitable for such an application.

The standard approach for sound analysis is to collect a single vector of features. The often used classifiers are Support Vector Machines (SVM), Deep Neural Networks (DNN), or multilayer perceptrons. When processing environmental sounds, the uniform structure can not be expected as in speech or music, where the signal contains a harmonic structure or repetitions. The features that perform well in specific applications may be inadequate for sounds with other nature and vice versa.

In our work we are using several 2-dimensional sound representations such as spectrograms as well as the standard 1-D approach and analyse the performance in the gunshot detection application. In addition to the frequently used spectrogram [12], [13], [14], also new visualizations and advanced methods for feature processing are used. The main contributions of our work are:

- Exploring the suitability of several state-of-the-art convolutional networks based approaches for gunshot detection.
- Proposing the convolutional model for boosting the performance on 2-dimensional independent feature spaces.

The signal is transformed into three independent audio feature sets forming an ”RGB image” that is suitable for processing by common 2-D convolutional networks used for image processing. Unlike image processing, where color channels are highly correlated and processed together after the first DNN layer, the three audio feature sets are processed independently by the first two DNN blocks. Our proposed architecture is based on standard residual units. The important task is not only increasing the number of true predictions but also reducing the number of false-positive gunshot predictions.

II. Related Work

Over the years several works dealing with gunshots detection from audio signal have been published. Most of them are based on the extraction of handcrafted acoustic features and the use of machine learning techniques for the task of classification. A combination of 7 Linear Predictive Coding (LPC) coefficients and 13 Mel Frequency Cepstral Coefficients (MFCC) with SVM classifier is used in work [15] providing 8 % false alarms rate on a custom dataset. The author of [16] extended the previous feature set by Linear Predictive Coding Cepstral (LPCC) and auto-correlation coefficients reaching 82 % accuracy and 70 % precision on a combination of public available datasets. The same author then evaluated the effect of individual features on the accuracy of the classification task [17], considering the features with the best score to be the first five coefficients of the 24th order LPC. A large set of various acoustic features with Hidden Markov Model (HMM) and Viterbi decoder is used in EAR-TUKE system [18] for detecting gunshots and glass breaking events with 98 % accuracy in records with Signal to Noise Ratio (SNR) $\geq$ 20 dB. In case of microphones arrays it is possible to use a two stage methodology comprising of a Blind System Identification and Deconvolution (BSID) stage followed by a SVM-based classification [19] for gunshot detection in a noisy urban environment. In [20] a method of classifying impulsive sounds based on a Weighted Majority Voting (WMV) strategy is described. In [21] Convolutional Neural Network (CNN) with temporal and spectral features is used for gunshot sound categories classification (pistol, rifle and shotgun of different calibres) reaching over 90 % accuracy. Another CNN approach detects gunshots with 99 % accuracy and low false alarm rate using the ResNet architecture [22]. Most of the above approaches work with database recordings that contain a low level of environmental noise. In the case of real applications, this condition can hardly be met. For this reason, dealing with the detection and classification of environmental noisy sounds is important.

The datasets of environmental sounds are made mainly for learning algorithms that perform Environmental Sound Classification (ESC) task. One of the widely used datasets for ESC is the UrbanSound8K [23], further described in Section III-E.

The signal representation of audio is related to the architecture of the learning algorithm and learning objective. The standard process that follows the approaches from speech and music analysis is to collect a single vector of features. Short-term or long-term features may not always generalize the unstructured nature of environmental sounds.

The direct solution of the feature extraction problem is to build a model that operates on the raw audio signals directly. The 1-D CNNs can handle the internal representation of the input signal, which allows end-to-end usage. The important advantage is that there is no need to transform or pre-process the data, and such a model can adapt to a variety of audio signals. In the studies [24], [25], the first end-to-end ESC architecture called EnvNet was proposed in versions 1 and 2. Another 1-D architecture was proposed in [26], where the audio signal was processed at different time scales. The study [27] presents an end-to-end 1-D convolutional network that has fewer parameters compared to dense 2-D convolutional neural networks and does not require a large amount of training data. It reaches 87 % mean accuracy on the UrbanSound8K dataset with random weights initialization and 89 % with weights initialization by Gammatone filter bank coefficients [28] synthesizing an impulse response from nerve cells in the auditory fiber [29].

The 2-D CNN models operate on the pre-computed feature representations obtained by a fixed process of extraction. The study [13] is the first which deals with ESR using CNN trained on mel-scaled spectrograms. Such an approach is extended in study [12], where different augmentation methods are used. In work [14], the authors presented the model ESResNet based on STFT, that outperforms recently known approaches with ESC datasets achieving 82% accuracy when trained from scratch and 85% accuracy with ImageNet weights initalization on UrbanSound8K dataset. Since the Piczak's work [13], the research trend in environmental sound analysis seems to be the usage of 2-D feature matrices for feeding 2-D CNNs [30].

III. Materials and Methods

A. Signal Model and Problem Formulation

Following the recent studies [12], [13], [14], [31], [32], [33], [34], the most suitable setup for environmental sound recognition employs a 2-D convolutional neural network fed by a Time-Frequency representation of the audio signal (further mentioned as audio features). In order to be processed by a 2-D convolutional network, these audio features need to be converted into a suitable uniform 2-D representation. Based on this 2-D representation, it is then necessary to choose an optimal architecture of the convolutional network for the classification task. The whole workflow from audio samples to predictions is depicted in Fig. 1.

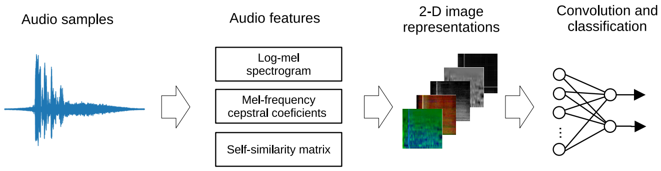

Audio samples
Audio features
Log-mel spectrogram
Mel-frequency cepstral coeficients
Self-similarity matrix
2-D image representations
Convolution and classification

Fig. 1. Feature Extraction and Classification Workflow.

B. Audio Features

In the case of audio signal processing, there is a large number of various audio features. The Log-Mel Spectrograms (LM Spec) and Mel Frequency Cepstral Coefficients (MFCC) are among the most commonly used audio features. In our work, we additionally include the Self-Similarity Matrix (SSM), which is frequently used to analyze the global structure of musical works, to the audio feature list. The hyperparameters for feature extraction are detailed in Table I.

1) Log-Mel Spectrogram: The spectrogram is the most commonly used audio signal visualization. It shows the frequency spectrum change over time. A Short Time Fourier Transform (STFT) is used to convert the signal from time to frequency domain. Additional mel-frequency scale transform 1 is applied to embrace the psychoacoustic knowledge.

$$f_{mel} = 2595 \cdot log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{1}$$

2) Mel Frequency Cepstral Coefficient: MFCC reflects the non-linear and masking psychoacoustic characteristics of human hearing. MFCC coefficients are obtained by multiplying the signal spectrum by a mel-scale distributed filter bank, logarithm and Discrete Cosine Transform (DCT).

3) Self-Similarity Matrix: SSM is the measure of self-similarity of the signal based on distances. We use a self-similarity matrix to display signal correlation. To visualize the self-similarity, we use a matrix S defined by Equation 2. The matrix dimensions $N \times N$ depend on the number of signal samples. For reducing the computational complexity, a self-similarity matrix S is computed on downsampled envelope $s = (s_1, s_2, s_3, ..., s_N)$ of input audio signal, obtained using the Hilbert transform. As a measure of similarity we are using the absolute distance.

$$S(i,j) = |s_i - s_j| \quad i,j = 1,...,N \tag{2}$$

The vertical and horizontal axes represent the time sequence. The matrix is symmetrical by the main diagonal where the similarity is maximal.

TABLE I. HYPERPARAMETERS FOR FEATURE EXTRACTION

Features	FFT length	Banks	Window
Log-mel spectrogram	2048	256	Hamming
MFCC	2048	20	Hamming
Self-similarity	-	-	-

C. 2-D Feature Representation

Audio features are extracted as 2-D matrices and aligned for convolutional neural network input. Since most 2D convolutional network architectures were primarily designed for image processing, they expect 3 sets of 2-D feature matrices at the input (an analogy to RGB channels of images).

1) Band-Splitted Spectrogram (BS Spec): In our experiments, we are using the log-mel spectrogram split to three frequency bands (high, middle, low) aligned with the RGB color channels (each band as one color channel). The band cutting frequencies depends the on maximal frequency $f_{max} = \frac{f_s}{2}$ given by the sampling rate (Fig. 2).

The same principle was used in study [14], where authors explain the usage of the band-splitted spectrogram for avoiding redundancy. Other solutions are replicating the spectrogram or passing zeros.

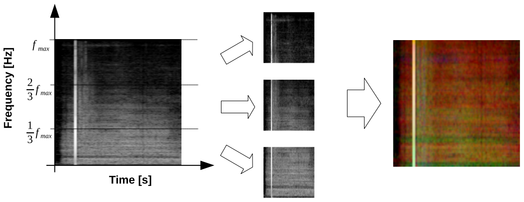
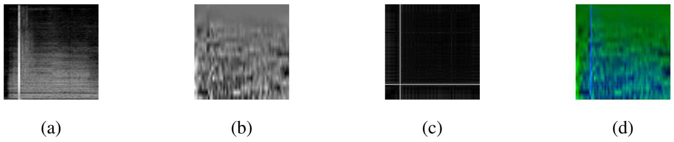


Fig. 2. Band-split Log-mel Spectrogram and Resulting RGB Image.

2) Independent Feature Spaces (IFS): The combination of the log-mel spectrogram, MFCC and self-similarity matrix represents the independent feature spaces. The similar method was used in study [32]. We assume that the MFCC and SSM will help to classify non-impulsive background sounds. The harmonicity of the gunshot signal is low, so the SSM is almost empty, while the background noise results in a visible grid.

Three feature matrices are overlapped in matched time positions. It means, that the x-axis resolutions are approximately the same for all matrices. However, on the y-axis we have different dimensions when using spectrogram (frequency), MFCC (mel banks) and SSM (time) (Fig. 3).



(a) (b) (c) (d)

Fig. 3. Independent Feature Spaces as RGB Image Channels. (a) Red Channel, Log-mel Spectrogram. (b) Green Channel, MFCCs. (c) Blue Channel, Self-similarity Matrix. (d) Resulting RGB Image.

D. Convolutional Neural Networks Architectures

The most widely used convolutional neural network models for 2-D feature space classification are based on the residual network architecture (ResNet). However, there is also an approach that uses only a 1-dimensional convolutional network fed directly by a raw audio signal to classify environmental sounds. In addition, we include a custom approach based on residual networks where individual channels are processed independently.

1) Residual Networks: Following recent studies [14], the residual models perform well on environmental sound classification. The residual network was designed as a network in a network, which means that the lower layer's inputs are connected to the outputs of the two higher layers. The example of the standard residual block is shown in Fig. 4.

The skip connections defined as

$$y = F(x) + x \tag{3}$$

are also called shortcut connections. The function F represents the convolution operations. The shortcut connections help to eliminate the problem of vanishing gradient in deep neural networks. The authors of [35] designed ResNets with a different number of layers, specifically 18, 34, 50, 101 and 152.

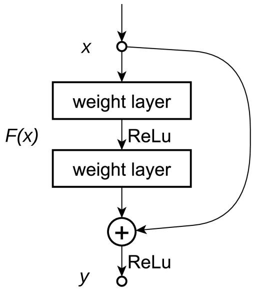


Fig. 4. Standard Residual Block [35].

2) End-to-End Classification using a 1-D CNN: In this approach, a 1-D convolutional (1-D CNN) neural network learns low-level and high-level information directly from the audio signal waveform. Since the size of the input data to the amount of data is in imbalance, it is not recommended to use too deep convolutional network architectures to avoid significant overfitting. The study [27] presents the optimal architecture with respect to the sampling frequency of the audio signal at different audio signal lengths. For the sampling frequency of 16 kHz considered further in this paper, it has been shown that the best score is achieved at the signal length of 1 second. The corresponding CNN architecture is shown in Table II consisting of 4 Convolutional Layers (CL), 2 Pooling Layers (PL) and 2 Fully Connected layers (FC). The Rectified Linear Unit (ReLU) activation function is used for all layers, except for the output layer where the softmax activation function is used with the output size equal to the number of classes beeing classified.

TABLE II. ARCHITECTURE OF 1-D CNN FOR 16 KHZ SAMPLING RATE AND AUDIO LENGTH OF 1 SECOND [27]

	CL1	PL1	CL2	PL2	CL3	CL4	FC1	FC2
Dimension	7969	996	483	60	23	8	128	64
Filters count	16	16	32	32	64	128	-	-
Filters size	64	8	32	8	16	8	-	-
Stride size	2	8	2	8	2	2	-	-

3) Independent Channel Residual Convolutional Network: We propose the Independent Channel Residual Convolutional Network (ICRCN), where the input RGB image is divided to the 2-D matrices in individual channels. The feature matrices share the dimension of the x-axis (time) but not the y-axis (frequency, mel banks, time). The system that combines different visual representations may suffer, when the features are combined as one input image. Therefore, we build the residual convolutional network, that processes different audio visualizations separately. The whole model architecture is shown in Table III.

The model input is a three channel RGB image. The separate channels contain residual blocks, where the number of filters is 32. The feature dimensions merging is made after the second residual block. From this point, the features are processed as in standard residual convolutional networks. The last convolutional block consists of 512 filters and it is followed by the classification layer. The proposed architecture is built up from standard residual blocks, as described in III-D1. The proposed architecture is compared to standard residual networks ResNet50 and 1-D CNN in Table IV.

TABLE III. PROPOSED INDEPENDENT CHANNEL RESIDUAL CONVOLUTIONAL NETWORK

Output size	ICRCN blocks		
(224, 224, 3)	Input RGB image		
(112, 112, 32)	7×7, 32	7×7, 32	7×7, 32
(56, 56, 32)	3x3, 32 3x3, 32 ×2	3x3, 32 3x3, 32 ×2	3x3, 32 3x3, 32 ×2
(28, 28, 64)	3x3, 64 3x3, 64 ×2	3x3, 64 3x3, 64 ×2	3x3, 64 3x3, 64 ×2
(28, 28, 192)	Concatenation		
(14, 14, 256)	3x3, 256 3x3, 256 ×2		
(7, 7, 512)	3x3, 512 3x3, 512 ×2		
(512)	Global average pooling		
(2)	Dense 2 + softmax		

TABLE IV. COMPARISON OF ARCHITECTURES COMPLEXITY

Architecture	Number of trainable parameters
1-D CNN	256k
ResNet50	23.5M
ICRCN	11M

E. Datasets

One of the most widely known datasets for environmental sound classification is the UrbanSound8K [23]. In this work we build our own gunshot detection dataset as a combination of the UrbanSound8k and The Free Firearm Sound Library [36].

- **UrbanSound8k** - 8732 tracks of 10 classes (air conditioner, car horn, children playing, dog barking, drilling, engine idling, gunshot, jackhammer, siren, street music), with varying sampling frequency file to file.
- **The Free Firearm Sound Library** - 2200 tracks of gunshots in a noise free environment including handguns (pistols, revolvers, semi-automatic pistols), rifles (lever-action, semi-automatic, fully automatic, machine guns, etc.) and shotguns. Recordings are in lossless wav format with sampling frequency 44.1 kHz.

The examples from The Free Firearm Sound Library were combined with gunshots from the UrbanSound dataset. The other sounds from UrbanSound were used as negative examples (random background). Since UrbanSound records are not equal in length, the window size of our examples floats in the range from 1 s to 4 s.

In the field of supervised machine learning, the performance of algorithms strongly depends on the quality of the dataset. To some extent, it is possible to simulate a big dataset using augmentation methods. However, this approach will never be as good as the expansion of the dataset by real examples. The following augmentation techniques are applied to positive examples (gunshots) while training.

- **Random background mixing** - The gunshot were randomly mixed with background noise in a random SNR from 0 dB to 20 dB.
- **Random time shift** - The onset positions of gunshots within the processing window were chosen randomly from 0 to 0.8 % of the window length.
- **Random Gaussian noise addition** - SNR from 60 dB to 100 dB.

F. Evaluation Metrics

The softmax output activation function directly indicates the probability of the example belonging to a certain class. The softmax activation also guarantees that the sum of probabilities over classes (background and gunshot) is one. The examples are classified as positive or negative according to higher or lower activation of output neurons. The confusion matrix can be constructed using predicted and true labels. Thereafter, the performance is evaluated via known metrics.

For practical implementation of the system for dangerous sound detection (gunshots, explosion, ...) it is necessary to minimize the amount of false-positive predictions. In this case, the high accuracy value may be a little bit confusing and thus we have to choose a metric for relevant evaluation. Therefore, in our work we use the following evaluation metrics:

- **Accuracy**: It reflects the overall algorithm performance, so it also takes into account the true negative predictions (TN). The accuracy is high when the number of true positive (TP) and true negative predictions is large. False positives (FP) and false negatives (FN) are prediction errors.

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \tag{4}$$

- **Sensitivity**: The relative amount of true positive predictions against all positive examples. It is often called *recall* or True Positive Rate (TPR).

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

- **Specificity**: The relative amount of true negative predictions against all negative examples.

$$Specificity = \frac{TN}{FP + TN} \tag{6}$$

- **False-Positive Rate (FPR)**: Reflects the relative amount of false positives against all negative examples.

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

- **F1 score**: The balance value between sensitivity and precision, where precision is the amount of true positives against all positive predictions. This means that the F1 score is not distorted by a large number of true negative predictions and may be considered as decisive, when testing data are unbalanced.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{8}$$

- **Detection Error Tradeoff** - DET curve visualizes the false positive rate vs. the false negative rate. Standardly, the axes are scaled non-linearly. Unlike the ROC curve, the DET curves are more linear and situated in most of the plot area.
- **Equal Error Rate** - EER is defined as the point in DET or ROC where the errors are equal. The lower EER value reflects better performance of the system.

IV. Results

All the networks were trained as long as validation loss was decreasing. Early stop was applied and only the best model was saved. Adam optimizer was used while training the models and the categorical cross-entropy loss was computed in each batch of 16 examples. Hardware used is NVIDIA GeForce GTX 1650, Intel Core i9-10900X CPU 3.7 GHz, RAM 64 GB. Fig. 5 shows the training and validation history of our ICRCN model when using 16 kHz sampling frequency.

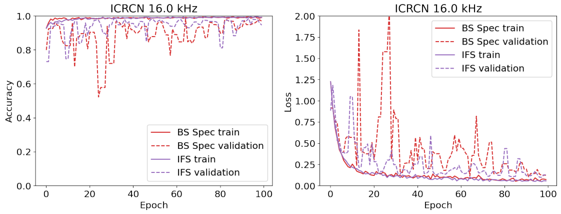


Fig. 5. Training and Validation History of ICRCN Model.

We use Tensorflow and Keras for building and training the models. For augmenting the audio we use the Audiomentations library. The Sci-kit learn is used for evaluation using described metrics.

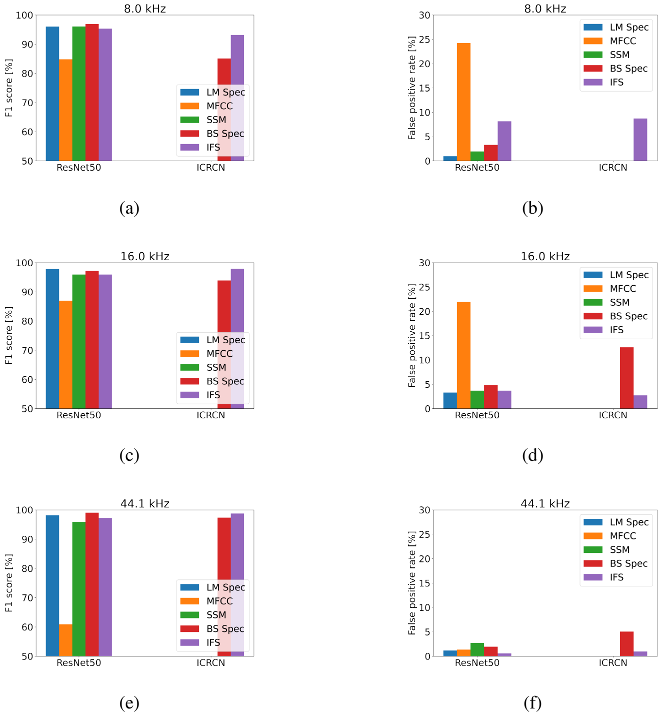
A. 2-D Convolutional Networks

The training audio examples were generated in three sampling frequencies to test the effect of sampling on system performance. The overall F1 score and FPR is evaluated in Fig. 6. As the test subset is balanced, the F1 score and overall accuracy metrics are very similar. As seen, the high F1 score value does not necessarily guarantee good performance in the reduction of false-positives.

When looking at the F1 score of individual models, the difference between 44.1 kHz and 16 kHz is not significant. In several cases, the score drops significantly at the 8 kHz sampling. The significant drop of F1 score is seen in case of ICRCN trained on band-split spectrograms. Table V, Table VI, and Table VII show also the change in sensitivity and specificity over all combinations.

TABLE V. Testing Results for Sampling Frequency 8 kHz

Model	Features	F1 [%]	Sensitivity [%]	Specificity [%]
ICRCN	IFS	93.1	94.8	91.3
	BS Spec	85.1	74.0	100.0
ResNet50	IFS	95.3	98.4	91.9
	LM Spec	96.0	93.2	99.0
	BS Spec	**96.9**	97.1	96.7
	MFCC	84.8	91.5	75.8
	SSM	96.0	94.2	98.1



(a) (b) (c) (d) (e) (f)

Fig. 6. Results for Different Sampling Frequencies. (a) F1 Score, 8 kHz. (b) FPR, 8 kHz. (c) F1 Score, 16 kHz. (d) FPR, 16 kHz. (e) F1 Score, 44.1 kHz. (f) FPR, 44.1 kHz.

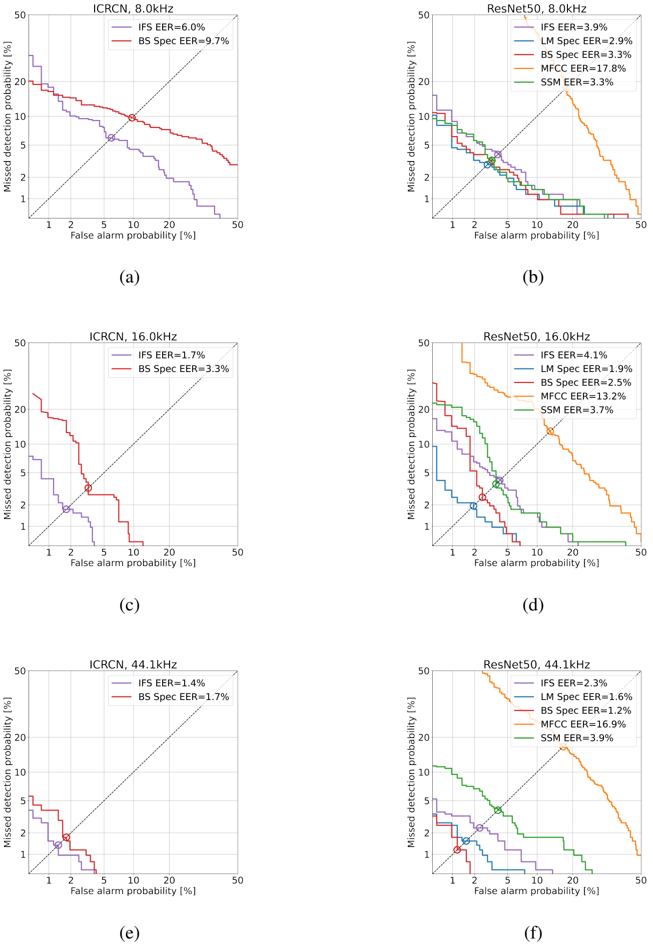
TABLE VI. Testing Results for Sampling Frequency 16 kHz

Model	Features	F1 [%]	Sensitivity [%]	Specificity [%]
ICRCN	IFS	**97.9**	98.4	97.3
	BS Spec	93.9	99.6	87.4
ResNet50	IFS	95.9	95.5	96.3
	LM Spec	97.8	98.8	96.7
	BS Spec	97.1	99.0	95.2
	MFCC	87.0	93.8	78.1
	SSM	95.9	95.5	96.3

TABLE VII. Testing Results for Sampling Frequency 44.1 kHz

Model	Features	F1 [%]	Sensitivity [%]	Specificity [%]
ICRCN	IFS	98.7	98.4	99.0
	BS Spec	97.3	99.6	95.0
ResNet50	IFS	97.2	95.2	99.4
	LM Spec	98.1	97.5	98.8
	BS Spec	**99.0**	100.0	98.1
	MFCC	60.9	44.4	98.6
	SSM	95.9	94.6	97.3

As the softmax output activation function is used, the output gives the probabilities of the example belonging to a certain class. Neverthelss, the operating point can be moved by changing the threshold of positive predictions. Choosing the operating point is strongly related to the application of the developed system and there are different ways how to set up the operating point. The useful visualizations are the DET curves (Fig. 7), which tend to be more linear and highlight the differences in the operating region more clearly than the ROC curves. The EER value represents the operating point, where the system performs equal in terms of the false-positive rate and false-negative rate. The false-positive rate can be interpreted as the false alarm probability and false-negative rate is the probability of missing the positive detection.



(a) (b) (c) (d) (e) (f)

Fig. 7. Detection Error Tradeoff Curves for Different Models and Sampling Frequencies. (a) ICRCN, 8 kHz. (b) ResNet50, 8 kHz. (c) ICRCN, 16 kHz. (d) ResNet50, 16 kHz. (e) ICRCN, 44.1 kHz. (f) ResNet50, 44.1 kHz.

As seen in Table 7, the EER evaluation slightly correlates with the overall F1 score of the system. Though there is no evident regress over the parameters, we can highlight several findings from the results. Using the original sampling frequency when the data are clear and detailed, all architectures model the data well. The more trainable parameters seem to be useful as the system losses the information by audio downsampling. At 16 kHz sampling, the performance of architectures is closely comparable. The proposed ICRCN performs best with IFS feature combination. According to study [32], the single SSM performs worst on environmental sounds classification. Our results showed, that for gunshot detection, using the SSM only may leads to better scores than single MFCC, which surprisingly performs worst.

B. 1-D Convolutional Network

In this part of work, we compare the 1-D CNN model to the best 2-D models. For this comparison we consider only the 16 kHz sampling frequency. As the best candidates from the previous test we choose two model-features combinations, namely ICRCN-IFS and ResNet50-LM Spec. The comparison is shown in Table VIII. The 1-D CNN model was trained on the same dataset as 2-D models.

TABLE VIII. The Results for 1-D CNN Model in Comparison to the Best 2-D Models

Model	Features	F1 [%]	Sensitivity [%]	Specificity [%]
ICRCN	IFS	97.9	98.4	97.3
ResNet50	LM Spec	97.8	98.8	96.7
1-D CNN	-	98.1	98.2	98.1

C. Evaluation of Performance in the Wild

The results on artificially generated test subsets may not always meet the real performance in the wild. Therefore, we evaluate the models on a real audio track from action movie to simulate the real world conditions. On this stage we use only 16 kHz sampling frequency as the most suitable based on previous results. The audio track from the movie John Wick (2014) was manually annotated. We use a 4 seconds window length with 2 seconds overlap and third-order median filtering of the CNN output. Each segment is labeled as positive if there is an occurrence of a gunshot. The numbers of positive and negative segments is very unbalanced, therefore we use the F1 score as the evaluation metric. The evaluation results are shown in Table IX.

TABLE IX. Evaluation on Real Data in the Wild

Model	Features	Test	Evaluation		
		F1[%]	F1[%]	Sensitivity[%]	Specificity[%]
ICRCN	BS Spec	93.9	14.5	98.2	10.5
	IFS	97.9	**44.3**	59.2	91.6
ResNet50	LM Spec	97.8	26.4	89.0	62.5
	MFCC	87.0	16.5	47.2	67.0
	SSM	95.9	5.5	5.5	92.7
	BS Spec	97.1	17.6	98.2	29.2
	IFS	95.9	16.2	95.4	24.3
1-D CNN	-	98.1	4.2	7.8	96.5

As shown in Table IX, the performance on real data drops significantly. However, the evaluation test scores correlate in a relative way. The test score difference is slight in most cases, while the evaluation score difference is noticeable. We can see, that single SSM outperforms the MFCC features in test results, but fails in evaluation, where the sensitivity and specificity are very unbalanced. The results show, that the proposed ICRCN network architecture was able to model the in a data robust way and outperforms the standard residual network ResNet50 even using less trainable parameters. In the case of 1-D CNN, the bad final score was probably caused by the low number of parameters of the convolutional neural network. Thus the network was not able to generalize the extracted features sufficiently to deal with the high variability of real audio data.

The previous experiment mainly focused on the ability of the proposed algorithm to detect gunshot sounds in the simulated real recording. However, in the case of surveillance systems, in addition to accuracy, a low frequency of false alarms is also required. This value cannot be derived from the above experiment because the soundtracks of the movies are sound-exposed. Thus, in this case, recordings from the QUT-NOISE database [37], which is the only suitable publicly available database containing continuous recordings of city sounds, were used to determine the false error rate. Since the total recording time is only a few hours, this database was supplemented with custom recordings from a busy city street. Within these recordings, the proposed algorithm did not detect a single false gunshot event.

D. Processing Time

In Table X, we showed the mean processing time for each of the features used in the experiments. The generation of three feature matrices takes twice as much time than generation of a single spectrogram. This must be taken into account if the detector is implemented on the system with limited processing capacity.

TABLE X. Feature Extraction Time

Sampling frequency [kHz]	Features	Time [ms]
8	Spec	5.02
	MFCC	2.68
	SSM	2.31
16	Spec	6.23
	MFCC	3.62
	SSM	3.71
44.1	Spec	10.53
	MFCC	6.9
	SSM	10.58

When comparing processing time between multiple sampling frequencies, the good compromise is to prefer the 16 kHz sampling. It offers the comparable performance to original sampling while the feature extraction time is almost halved.

V. Discussion

When comparing the residual architectures and features using testing data, there is not a significant best setup. The evaluation on real data showed the benefit of independent features processing, where the difference in score is observable. The standard ResNet50 performs well with a single spectrogram. The feature combination works best with the proposed architecture with independent channels. Therefore, we prefer using the proposed ICRCN architecture and IFS for gunshot detection. Taking the feature extraction times into account, we prefer using the 16 kHz sampling.

We assume, that the ICRCN model fed by IFS matrices is the most suitable for gunshot detection systems, outperforming the other state-of-the-art approaches (in terms of F1 score). The combination of spectrogram, MFCC, and SSM was used in study [32], with conclusion that it does not improve the accuracy of environmental sounds classification. We assume, that the feature combination can be beneficial for gunshot detection applications and has a potential to lower the false-positive rate. The results showed, that there is a benefit of splitting the convolutional channels. The real performance of the system relates on the operating point. In some applications the specificity may be a more important score than sensitivity, or vice versa. We visualized the DET curves, as they can help to fit the requirements of the system for particular applications.

VI. Conclusion

In this paper, we analyzed the usage of several convolutional architectures for the gunshot detection task. We proposed a new architecture and compared its performance to standard ResNet50 and 1-D architectures. The effect of different features on the resulting performance was tested using several Time-Frequency audio representations. For training, validation and testing we collected the gunshots and random backgrounds audio clips from public datasets. In the training phase, the standard augmentation methods were used. Finally, we simulated the real world conditions by evaluation of a real audio track from the action movie.

We achieved the goal of our work by proposing a system for gunshot audio events detection. Our ICRCN approach is able to operate in noisy environments with high specificity (slightly over 90 %), by maintaining fair sensitivity (almost 60 %). The Detection Error Tradeoff analysis showed, that the real performance of the system strongly depends on the error tolerance and requirements. The operating point should be selected for specific applications. On our test data, the missed detection probability is about 10 % when the false alarm probability is as minimal as possible. We expect the similar results also for explosion detection. Such a systems can be implemented in a complex application together with a smoke or fire detector. Due to the fact that the proposed approach has a relatively low computational time, it can be easily integrated into existing surveillance systems without the need to invest in expensive computing servers to operate in real time.

Acknowledgment

This work was supported by the Slovak Grant Agency KEGA under contract no. KEGA 008ZU-4/2021.

References

[1] J. P. Bello, C. Silva, O. Nov, R. L. DuBois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A System for the Monitoring, Analysis and Mitigation of Urban Noise Pollution," *arXiv:1805.00889 [cs, eess]*, May 2018, arXiv: 1805.00889.

[2] S. C. Tan and A. Abd Manaf, "Characterization of internet of things (iot) powered-acoustics sensor for indoor surveillance sound classification," in *2021 IEEE International Conference on Sensors and Nanotechnology (SENNANO)*. IEEE, 2021, pp. 109–112.

[3] F. Graf and M. Gruber, "Rapid Incident Detection in Tunnels through Acoustic Monitoring – Operating Experiences in Austrian Road Tunnels," Graz, Jan. 2018, p. 8.

[4] C. Watkins, L. Green Mazerolle, D. Rogan, and J. Frank, "Technological approaches to controlling random gunfire: Results of a gunshot detection system field test," *Policing: An International Journal of Police Strategies & Management*, vol. 25, no. 2, pp. 345–370, Jun. 2002.

[5] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," Oct. 2007, pp. 21–26.

[6] M. Sigmund and M. Hrabina, "Efficient feature set developed for acoustic gunshot detection in open space," *Elektronika ir Elektrotechnika*, vol. 27, no. 4, pp. 62–68, 2021.

[7] H. Luo, J. Liu, W. Fang, P. E. Love, Q. Yu, and Z. Lu, "Real-time smart video surveillance to manage safety: A case study of a transport mega-project," *Advanced Engineering Informatics*, vol. 45, p. 101100, 2020.

[8] N. Khalid, M. Gochoo, A. Jalal, and K. Kim, "Modeling two-person segmentation and locomotion for stereoscopic action identification: A sustainable video surveillance system," *Sustainability*, vol. 13, no. 2, 2021.

[9] P. K.-Y. Wong, H. Luo, M. Wang, P. H. Leung, and J. C. Cheng, "Recognition of pedestrian trajectories and attributes with computer vision and deep learning techniques," *Advanced Engineering Informatics*, vol. 49, p. 101356, 2021.

[10] W. Renda and C. H. Zhang, "Comparative analysis of firearm discharge recorded by gunshot detection technology and calls for service in louisville, kentucky," *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, p. 275, 2019.

[11] J. H. Ratcliffe, M. Lattanzio, G. Kikuchi, and K. Thomas, "A partially randomized field experiment on the effect of an acoustic gunshot detection system on police incident reports," *Journal of Experimental Criminology*, vol. 15, no. 1, pp. 67–76, 2019.

[12] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, Mar. 2017, arXiv: 1608.04363.

[13] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Boston, MA, USA: IEEE, Sep. 2015, pp. 1–6.

[14] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "ESResNet: Environmental Sound Classification Based on Visual Domain Models," *arXiv:2004.07301 [cs, eess]*, Apr. 2020, arXiv: 2004.07301.

[15] T. Ahmed, M. Uppal, and A. Muhammad, "Improving efficiency and reliability of gunshot detection systems," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 513–517.

[16] M. Hrabina and M. Sigmund, "Gunshot recognition using low level features in the time domain," in *2018 28th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2018, pp. 1–5.

[17] M. Hrabina, "Analysis of linear predictive coefficients for gunshot detection based on neural networks," in *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*. IEEE, 2017, pp. 1961–1965.

[18] M. Lojka, M. Pleva, E. Kiktová, J. Juhár, and A. Čižmár, "Efficient acoustic detector of gunshots and glass breaking," *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 10 441–10 469, 2016.

[19] A. A. Shiekh, M. Tahir, and M. Uppal, "Accurate gunshot detection in urban environments using blind deconvolution," in *2017 International Multi-topic Conference (INMIC)*. IEEE, 2017, pp. 1–4.

[20] A. Suliman, B. Omarov, and Z. Dosbayev, "Detection of impulsive sounds in stream of audio signals," in *2020 8th International Conference on Information Technology and Multimedia (ICIMU)*. IEEE, 2020, pp. 283–287.

[21] S. Raponi, I. Ali, and G. Oligeri, "Sound of guns: digital forensics of gun audio samples meets artificial intelligence," *arXiv preprint arXiv:2004.07948*, 2020.

[22] J. Bajzik, J. Prinosil, and D. Koniar, "Gunshot detection using convolutional neural networks," in *2020 24th International Conference Electronics*. IEEE, 2020, pp. 1–5.

[23] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the ACM International Conference on Multimedia - MM '14*. Orlando, Florida, USA: ACM Press, 2014, pp. 1041–1044.

[24] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 2721–2725.

[25] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from Between-class Examples for Deep Sound Recognition," *arXiv:1711.10282 [cs, eess, stat]*, Feb. 2018, arXiv: 1711.10282.

[26] B. Zhu, K. Xu, D. Wang, L. Zhang, B. Li, and Y. Peng, "Environmental Sound Classification Based on Multi-temporal Resolution Convolutional Neural Network Combining with Multi-level Features," *arXiv:1805.09752 [cs, eess]*, Jun. 2018, arXiv: 1805.09752.

[27] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-End Environmental Sound Classification using a 1D Convolutional Neural Network," *arXiv:1904.08990 [cs, stat]*, Apr. 2019, arXiv: 1904.08990.

[28] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, "Practical gammatone-like filters for auditory processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. 1–15, 2007.

[29] H. Park and C. D. Yoo, "Cnn-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification," *IEEE Signal Processing Letters*, vol. 27, pp. 411–415, 2020.

[30] I. Papadimitriou, A. Vafeiadis, A. Lalas, K. Votis, and D. Tzovaras, "Audio-based event detection at different snr settings using two-dimensional spectrogram magnitude representations," *Electronics*, vol. 9, no. 10, p. 1593, 2020.

[31] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Detection," *arXiv:1604.07160 [cs]*, Dec. 2016, arXiv: 1604.07160.

[32] 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.

[33] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep Convolutional Neural Network with Mixup for Environmental Sound Classification," *arXiv:1808.08405 [cs, eess]*, Aug. 2018, arXiv: 1808.08405.

[34] H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental Sound Classification with Parallel Temporal-spectral Attention," *arXiv:1912.06808 [cs, eess]*, May 2020, arXiv: 1912.06808.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385.

[36] "Sound Effects Library," Mar. 2013.

[37] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "Gthe qut-noise-timit corpus for evaluation of voice activity detection algorithms," in *2010 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 3110–3113.