STFT and Gaussian Filter-Based Soundscape classification via 2D Convolutional Neural Networks

Razika SOUADEK[(1)*]
[(1)] Mechanical and Electromechanical Engineering Department, Institute of Sciences and Technology, University of Mila, The Environmental Engineering and Technologies Laboratory (EE&TLab), Mila, Algeria
[*] r.souadek@centre-univ-mila.dz

Abstract: A novel approach is presented for soundscape classification, relying on STFT-based spectral characteristics combined with Gaussian filtering to generate enhanced time–frequency representations. Data augmentation is achieved through pitch shifting and time stretching to increase variability in the audio sequences. The signals are transformed using the Short-Time Fourier Transform and subsequently converted to the decibel scale. A normalized Gaussian filter is then applied to yield the Gaussian-STFT feature maps. For the classification stage, a two-dimensional convolutional neural network composed of three convolutional blocks followed by dense layers is employed. Experimental evaluation on the ESC-10 dataset achieves an accuracy of 97%, a precision of 97.34%, a recall of 97.08%, and an F1-score of 97.11%. Evaluation on the Urban Sound dataset yields an accuracy of 90%, a precision of 90.59%, a recall of 89.96%, and an F1-score of 89.97%. Evaluation on the ESC-50 dataset results in an accuracy of 90%, a precision of 90.71%, a recall of 90.00%, and an F1-score of 89.94%.

Keywords: STFT, CNN, Sound Classification.

1. INTRODUCTION

The urban acoustic scene contains of a wide variety of sounds, such as those produced by humans and their activities, animal sounds, various modes of transportation, sirens, alarms, and industrial noise [1]. This diversity makes the analysis and classification of sounds in an urban environment particularly complex. Identifying, organizing, and isolating these sounds represents a significant challenge for researchers, mainly when it comes to detecting abnormal events in real time [2].

Sound protection now corresponds to a broader aspect of the evolution of smart cities, where public safety and quality of life partly depend on effective management of the sound environment [3]. The prevention of nuisances and the automatic detection of critical sounds (such as accidents or assaults) have become priorities in many connected urban projects.

Sound classification was commonly based on statistical Gaussian mixture representations [4]. However, the evolution of deep learning has enabled the emergence of more effective techniques, capable of automatically learning complex acoustic representations from large amounts of data [5].

In this study, we present an algorithm for urban sound classification based on data augmentation through pitch shifting and time stretching. Audio features are extracted using the Short-Time Fourier Transform (STFT), then normalized using a Gaussian divergence function. Finally, a lightweight convolutional model, structured in three layers with MaxPooling operations and a Softmax output layer, is used to perform the final classification.

2. RELATED WORK

The researchers in [6] propose a CRNN (Convolutional Recurrent Neural Network) as a classification model and use raw waveforms of acoustic characteristics. The extracted sound features are used as input to the CRNN for temporal aggregation. The experimental results show an accuracy of 79.06%.

The authors [7] proposed a novel method based on Deep Convolutional Neural Networks (CNNs) in this paper. The audio signals are first denoised and then transformed into spectrogram images using Short-Time Fourier Transform (STFT), which are used as inputs to the CNN models. A Support Vector Machine (SVM) classifier is applied at the final stage. The method achieved accuracy scores of 94.8%, 81.4%,

and 78.14% on the ESC-10, ESC-50, and UrbanSound8K datasets, respectively.

This research uses a DCNN structured into two models: one with max-pooling (Model-1) and one without max-pooling (Model-2). The feature extraction is performed using Mel spectrogram (Mel), Mel Frequency Cepstral Coefficient (MFCC), and Log-Mel. Model-2 achieves the best accuracy, with 94.94% for ESC-10, 89.28% for ESC-50, and 95.37% for UrbanSound8K [8].

The study reached an algorithm based on the DNN model performance and compared it with the CNN and LSTM models' performance. The characteristic used in classifying urban sound is spectrum images of Mel scale cepstral (MEL). The LSTM model shows good accuracy performance on test data with 90.15% [9].

3. TWO-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK

The two-dimensional convolutional neural network is a deep model used to process images and audio representations in the form of spectrograms. It uses convolutional filters to extract the desired features. These filters are then slid across the input to compute convolutions. ReLU activation is applied afterward, followed by MaxPooling to reduce redundant information and retain only the most relevant features.

4. GAUSSIAN-STFT FEATURES

In this feature extraction phase, we rely on the Short-Time Fourier Transform (STFT), followed by the application of the Gaussian filter (Fig. 1). All components involved in extracting these features are detailed as follows:

Short Time Fourier Transform

First, this transform divides the signal into fixed-length windows, and then applies the Fourier transform to each window separately to obtain the STFT spectrum.

$$STFT\{x[n]\} \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]\omega[n-m]e^{-j\omega n}$$

$x[n]$ is the discrete signal.

$\omega[n-m]$ is the windowing function.

The value absolute of Short Time Fourier Transform is:

$$STFT_{abs} \equiv |X(m,\omega)|$$

Conversion of the absolute value of STFT in decibel is:

$$STFT_{dB} = 10\log_{10}\left(\frac{STFT_{abs}{}^2}{\max(|X|)^2} + \epsilon\right)$$

ϵ is a small constant to avoid division by zero.

Normalized Filter Gaussian

To obtain a blurred image inspired by the centered Gaussian distribution.

$$G(m,\omega) = \exp\left(-\frac{1}{2}\left(\frac{STFT_{dB}-\mu}{\sigma+\epsilon}\right)^2\right)$$

$$\mu = mean(STFT_{dB})$$

$$\sigma = std(STFT_{dB})$$

This transforms the intensity values into a centered Gaussian function, creating a blurred map where intensities close to the mean are maximized.

5. PROPOSED ALGORITHM

The proposed urban sound classification algorithm presented in this paper is based on a two-dimensional convolutional neural network (2D CNN), the Short-Time Fourier Transform (STFT), and Gaussian filtering (Fig. 2). The algorithm is summarized as follows:

- Load audio files from the dataset.
- Apply data augmentation: pitch shifting and time stretching.
- Compute the STFT for each audio signal (including its augmented versions).
- Convert the spectrum to decibel scale, then apply Gaussian normalization.
- Resize the spectrogram image to the format 128×128×3.
- Use a 2D CNN model composed of several blocks:

CNN Model Structure

Three convolutional blocks:

each consisting of:

- A Conv2D layer with (3×3) filters and ReLU activation
- A MaxPooling2D (2×2) layer to reduce spatial dimensions
- A Dropout layer to prevent overfitting

Final layers

- A Flatten layer to convert feature maps into a vector
- A dense layer with 256 neurons and ReLU activation
- A Dropout layer with a 30% rate
- A Softmax output layer to produce class probability scores

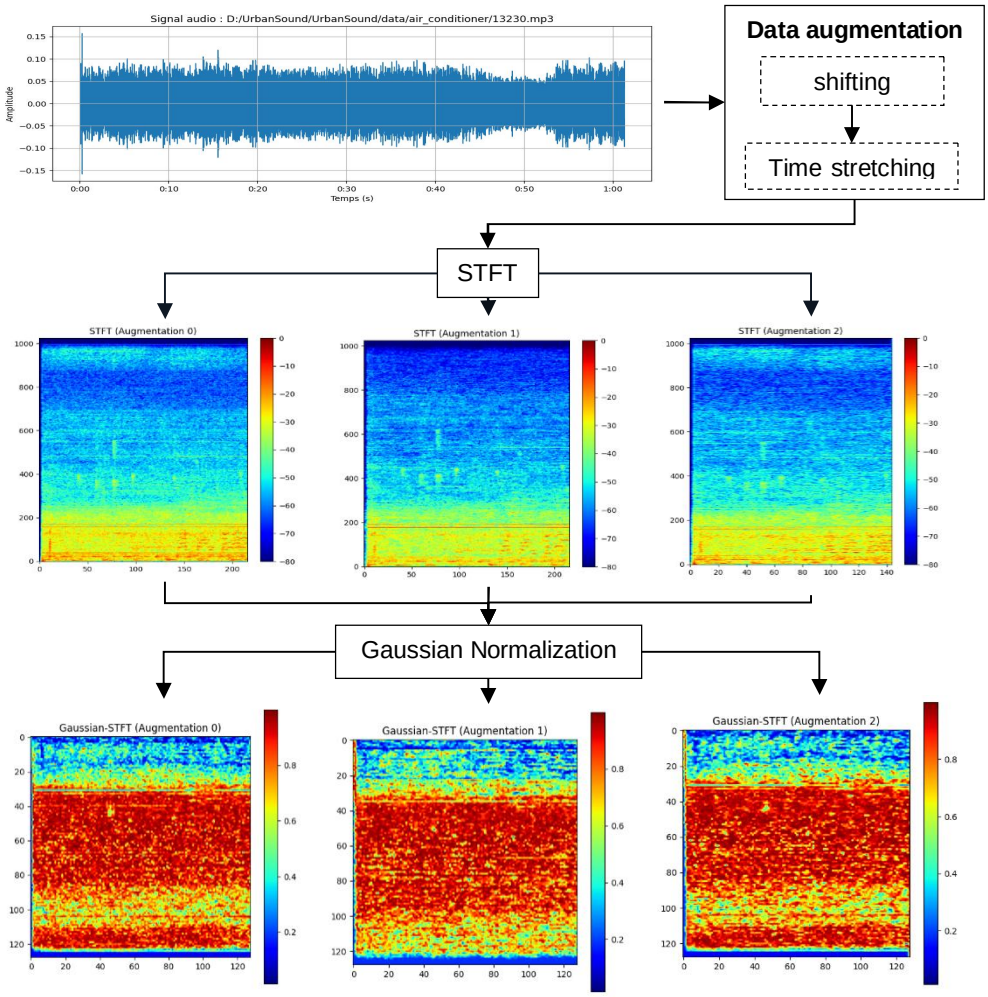


Fig. 1: Gaussian-STFT Features

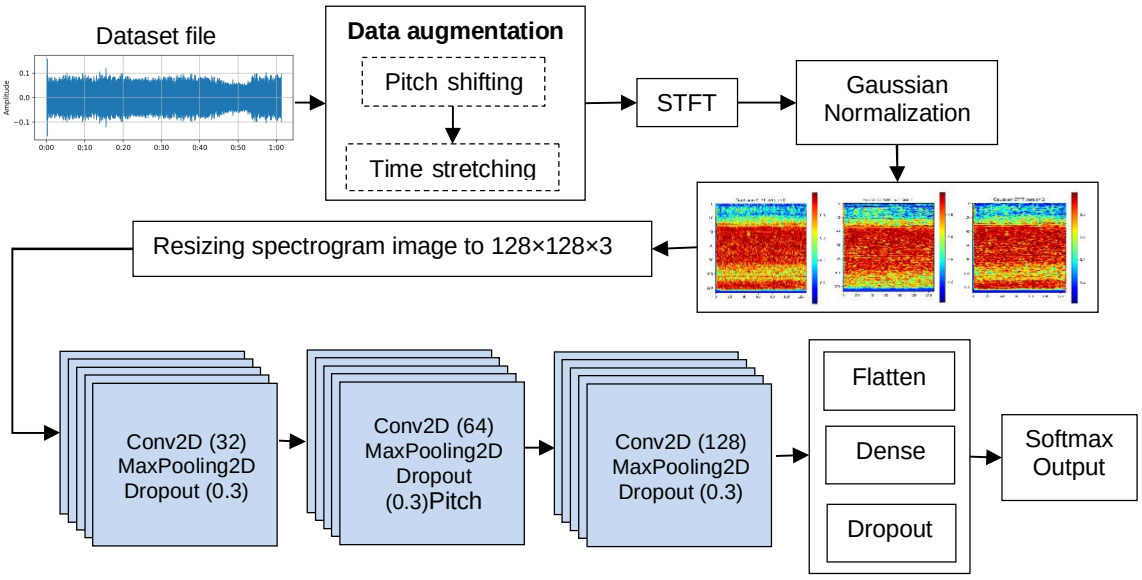


Fig. 2: Proposed Urban Sound Classification Algorithm

6. EXPERIMENTAL RESULT

In this section of the results we present all the obtained values of accuracy and loss for each dataset (Fig.3).

Dataset: The datasets used in this part of the experiments are UrbanSound, consisting of 1,302 audio samples [10]; ESC-50, containing 2,000 audio recordings [11]; and ESC-10, comprising 400 sound scenes [12].

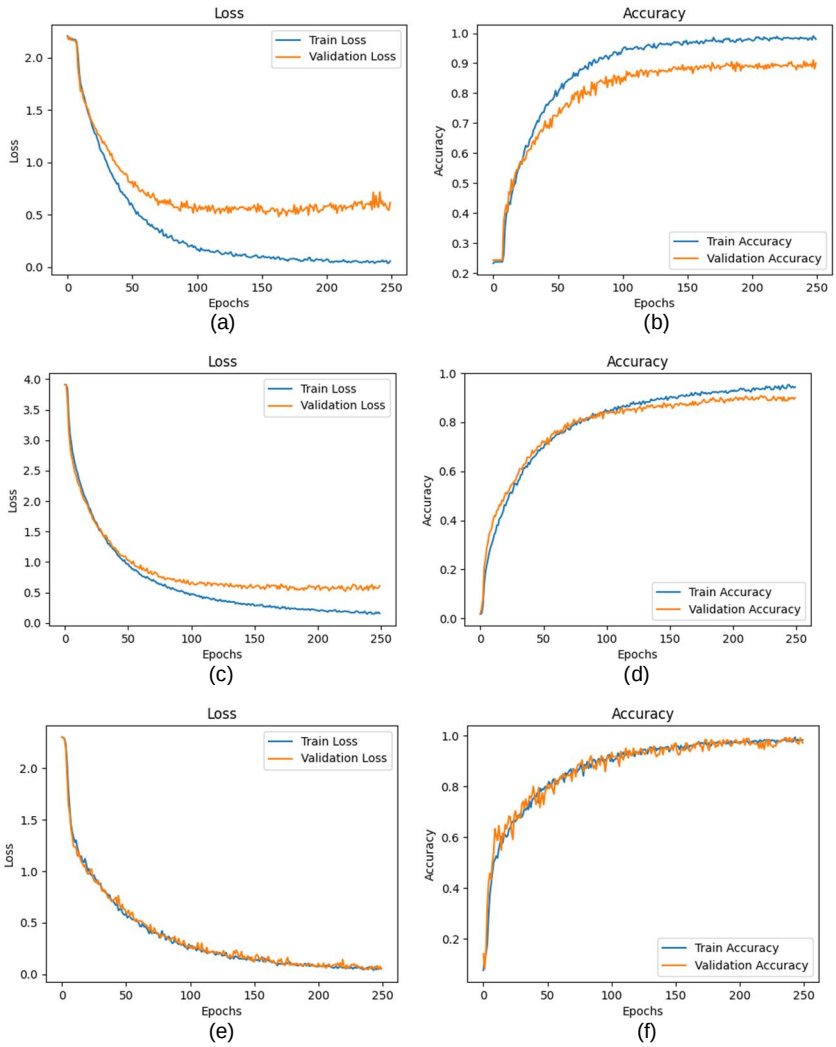


Fig.3: urban sound classification based on Gaussian-STFT features and CNN. (a) loss curve for UrbanSoundK dataset (b) accuracy curve for UrbanSoundK dataset, (c) loss curve for ESC 50 dataset (d) accuracy curve for ESC 50 dataset, (e) loss curve for ESC 10 dataset,(f) accuracy curve for ESC 10 dataset.

In the figure 1, part (a) (UrbanSound Dataset), the loss decreases steadily and continuously, unlike part (b), which shows the accuracy curve with a noticeable increase that stabilizes around epoch 150 at approximately 93%. In the same figure, part (c) (ESC 50 Dataset) also shows a decreasing loss curve, while part (d), the accuracy curve, increases consistently, reaching nearly 87%, indicating good performance.

The training and validation loss decrease steadily and converge to low values (e), while the accuracy curves increase rapidly and remain very close to each other (f). This behavior indicates fast convergence, good generalization, and no overfitting. The high final accuracy confirms the effectiveness of the proposed Gaussian-STFT and 2D CNN model on the ESC-10 dataset.

Table 1: Results of metrics (precision, recall, F1 score, accuracy) for UrbanSound dataset.

Class	Precision	recall	F1-score	Accuracy
air_conditioner	0.84	0.8	0.82	0.8
car_horn	0.79	0.94	0.86	0.94
children_playing	0.95	0.93	0.94	0.93
dog_bark	0.94	0.99	0.96	0.99
drilling	0.93	0.77	0.84	0.7
engine_idling	0.91	0.79	0.85	0.78
gun_shot	0.98	0.93	0.96	0.93
jackhammer	0.67	0.76	0.71	0.76
siren	0.8	0.9	0.85	0.90
street_music	0.98	0.84	0.9	0.83
Average	0.90	0.89	0.89	0.90

The Table 1 shows excellent performance on the UrbanSound dataset, with high scores across all metrics.

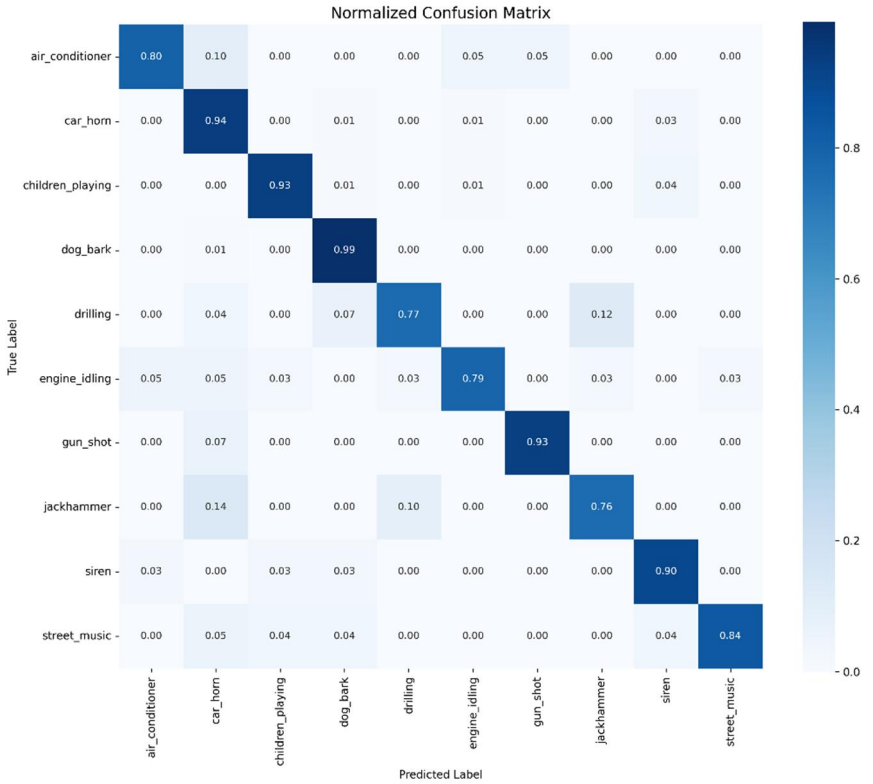


Fig. 4: Normalized Confusion Matrix for UrbanSound dataset

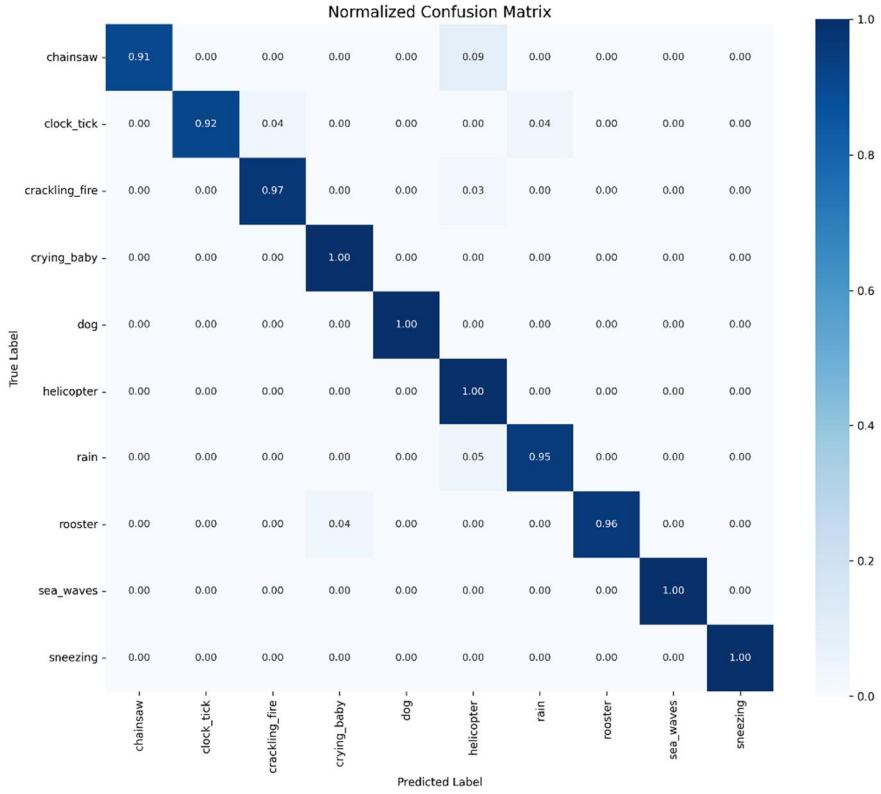


Fig. 5: Normalized Confusion Matrix for ESC 10 dataset

The class 7 demonstrates that the model performs very well when the audio features are clear and distinctive. The Fig.4 and Fig.5 is presented overall accuracy is very good, with the majority of classes well distinguished for Urbansound and ESC 10 datasets.

Table 2: Results of metrics (precision, recall, F1 score, accuracy) for ESC 50 dataset.

Classes	precision	recall	F1 score	Accuracy
101-Dog	0.97	1	0.98	1
102 – Rooster	0.96	0.96	0.96	0.96
103 – Pig	0.81	0.97	0.88	0.96
104 – Cow	1	0.88	0.93	0.87
105 – Frog	0.89	1	0.94	1
106 – Cat	0.95	0.86	0.9	0.85
107 – Hen	0.95	0.68	0.79	0.67
108 – Insects	0.94	1	0.97	1
109 – Sheep	0.87	1	0.93	1
110 – Crow	0.85	1	0.92	1
201 – Rain	0.9	0.86	0.88	0.85
202 – Sea waves	1	0.96	0.98	0.96
203 – Crackling fire	0.97	0.97	0.97	0.96
204 – Crickets	0.89	0.93	0.91	0.92
205 – Chirping birds	0.96	0.92	0.94	0.92
206 – Water drops	1	0.78	0.88	0.77
207 – Wind	0.77	0.89	0.83	0.89
208 – Pouring water	1	1	1	1
209 – Toilet flush	1	1	1	1
210 – Thunderstorm	0.89	1	0.94	1
301 – Crying baby	1	0.75	0.86	0.75
302 – Sneezing	0.91	0.95	0.93	0.95
303 – Clapping	0.94	1	0.97	1
304 – Breathing	0.86	0.86	0.86	0.85
305 – Coughing	0.93	0.81	0.87	0.81
306 – Footsteps	0.81	0.96	0.88	0.96
307 – Laughing	0.92	0.88	0.9	0.88
308 – Brushing teeth	0.88	0.88	0.88	0.88
309 – Snoring	0.81	0.89	0.85	0.89
310 – Drinking – sipping	1	0.96	0.98	0.96
401 – Door knock	1	0.9	0.95	0.90
402 – Mouse click	0.87	0.93	0.9	0.93
403 – Keyboard typing	0.96	0.93	0.94	0.92
404 – Door – wood creaks	0.74	1	0.85	1
405 – Can opening	0.93	0.82	0.88	0.82
406 – Washing machine	0.7	0.83	0.76	0.82
407 – Vacuum cleaner	0.9	0.75	0.82	0.75
408 – Clock alarm	0.87	0.91	0.89	0.91
409 – Clock tick	0.92	0.74	0.82	0.74
410 – Glass breaking	0.86	1	0.93	1
501 – Helicopter	0.83	0.91	0.87	0.91
502 – Chainsaw	0.96	0.88	0.92	0.88
503 – Siren	0.86	0.9	0.88	0.90
504 – Car horn	0.91	0.91	0.91	0.91
505 – Engine	0.6	0.57	0.59	0.57
506 – Train	0.88	0.88	0.88	0.88
507 – Church bells	1	0.72	0.84	0.72
508 – Airplane	0.97	1	0.98	1
509 – Fireworks	0.96	0.83	0.89	0.82
510 – Hand saw	0.71	1	0.83	1
Average	0.90	0.89	0.89	0.90

The model achieved solid average values across all metrics: Average Precision: 0.90, Average Recall: 0.89, Average F1-score: 0.89, Average Accuracy: 0.90. The obtained results demonstrate the effectiveness of proposed algorithm. This reflects a generally

balanced and reliable performance across all classes in the ESC-50 dataset (Table 2).

Table 3: Results of metrics (precision, recall, F1 score, accuracy) for ESC 10 dataset.

Class	Precision	Recall	F1-Score	Accuracy
chainsaw	1	0.91	0.95	0.90
clock_tick	1	0.92	0.96	0.91
crackling_fire	0.97	0.97	0.97	0.96
crying_baby	0.96	1	0.98	1
dog	1	1	1	1
helicopter	0.86	1	0.92	1
rain	0.95	0.95	0.95	0.95
rooster	1	0.96	0.98	0.96
sea_waves	1	1	1	1
sneezing	1	1	1	1
Average	0.97	0.97	0.97	0.97

The model achieves excellent and well-balanced performance across all ESC-10 classes, with most categories reaching near-perfect precision, recall, and F1-scores. The results confirming strong generalization and robustness (Table 3).

The obtained results determine the accuracy performance of proposed model compared with the other researches (Table 4).

Table 4: Comparison of accuracy results for ESC 50 dataset

Dataset	[7]	[8]	Proposed model
ESC 50	81.4%	89.2%	90 %

7. CONCLUSION

This paper presents an efficient urban sound classification approach based on Gaussian-STFT feature representations and a two-dimensional convolutional neural network (2D CNN). Extensive experiments conducted on three benchmark datasets demonstrate the effectiveness of the proposed method. Specifically, the model achieves an accuracy of 97% on the ESC-10 dataset, with precision, recall, and F1-score exceeding 97%. On the UrbanSound dataset, the proposed approach attains an accuracy of 90%, along with a precision of 90.59%, a recall of 89.96%, and an F1-score of 89.97%. Similarly, strong performance is observed on the ESC-50 dataset, with an accuracy of 90%, a precision of 90.71%, a recall of 90.00%, and an F1-score of 89.94%.

Compared with recent studies, these results highlight the advantages of combining Gaussian normalization with STFT-based representations and data augmentation techniques, leading to improved robustness and generalization. The proposed framework provides valuable insights into urban soundscape analysis and holds significant potential for real-world applications such as public safety monitoring and environmental noise management.

8. REFERENCES

[1] Zilong Huang, Chen Liu, Hongbo Fei, Wei Li, Jinghu Yu, Yi Cao, Urban sound classification based on 2-order dense convolutional network using dual features Applied Acoustics, Volume 164,2020,107243,

[2] DEMIR, Fatih, ABDULLAH, Daban Abdulsalam, et SENGUR, Abdulkadir. A new deep CNN model for environmental sound classification. IEEE Access, 2020, vol. 8, p. 66529-66537.

[3] NOGUEIRA, Ana Filipa Rodrigues, OLIVEIRA, Hugo S., MACHADO, Jose JM, et al. Sound classification and processing of urban environments: A systematic literature review. Sensors, 2022, vol. 22, no 22, p. 8608.

[4] Shervegar MV, Bhat GV. Heart sound classification using Gaussian mixture model. Porto Biomed J. 2018 Aug 15;3(1)

[5] İNIK, Özkan. CNN hyper-parameter optimization for environmental sound classification. Applied Acoustics, 2023, vol. 202, p. 109168.

[6] Jonghee Sang, Soomyung Park, Junwoo Lee, Convolutional Recurrent Neural Networks for Urban Sound Classification using Raw Waveforms, 2018 26th European Signal Processing Conference (EUSIPCO). 2458-2462.

[7] Fatih Demir, Muammer Turkoglu , Muzaffer Aslan, Abdulkadir Sengur, A new pyramidal concatenated CNN approach for environmental sound classification, Applied Acoustics, 170 (2020) 107520. 1-7.

[8] Zohaib Mushtaq, Shun-Feng Su, Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Applied Acoustics 167 (2020) 107389, 1-13.

[9] Mohamed Bubashait, Nabil Hewahi, " Urban Sound Classification Using DNN, CNN & LSTM a Comparative Approach" 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)

[10] https://urbansounddataset.weebly.com/urbansound.html

[11] Piczak, K.J. ESC-50: Dataset for Environmental Sound Classification. Available online: https://github.com/karolpiczak/ESC-50 (accessed on 8 July 2023).

[12] https://www.kaggle.com/datasets/sreyareddy15/esc10rearranged