EFFECTIVE SOUND CLASSIFICATION FOR NOISE FILTERING USING CNN WITH TENSOR WAVELET INTEGRATION

T Leela Rani, Student, Department of Computer Science and Engineering, LENDI INSTITUTE OF ENGINEERING AND TECHNOLOGY leelarani9817@gmail.com,
A Rama Rao, *PROFESSOR,* Department of Computer Science and Engineering, LENDI INSTITUTE OF ENGINEERING AND TECHNOLOGY
Golagani A V R C Rao, **S RamaKrishna**, Assoc.Prof, Department of Computer Science and Engineering, LENDI INSTITUTE OF ENGINEERING AND TECHNOLOGY

ABSTRACT Sound is essential to every aspect of human existence. Sound is an essential component when it comes to the development of automated systems for a wide range of applications, including personal security and vital surveillance. There are currently a few systems on the market, however the efficiency of these systems is a worry for applications in the real world. Deep learning architectures may be used to their full potential to learn, which can then be used to construct robust categorization systems that can overcome the inefficiencies of traditional methods. Deep learning networks are going to be utilised in this study to classify environmental sounds based on the spectrograms that are produced by these networks. When we were training the convolutional neural network (CNN) and the Tensor Wavelet , we made use of spectrograms that were taken of environmental noises. Kaggle common voice and common voice 2 were the datasets that were utilized for this inquiry. Both of these systems underwent training using the aforementioned datasets, with the CNN achieving an accuracy of 77% and the trained on the Common voice data set achieving an accuracy of 85%. Based on the results of this experiment, the method provided for sound classification that uses spectrogram images of sounds has the potential to be utilised effectively in the process of developing sound classification and recognition systems.

INDEX TERMS Deep learning, convolutional neural network, tensor deep stacking networks, spectrograms

I. INTRODUCTION:

In recent years, research on automatic sound recognition has helped a variety of fields, including multimedia [1], bioacoustics monitoring [2], intrusion detection [3] in wildlife regions, audio surveillance [4], and environmental sounds [5]. There are three distinct phases involved in the sound identification problem: signal preprocessing, feature extraction, and feature classification. By segmenting the input data, related features can be extracted during pre-processing. Data is simplified and complicated information is represented as feature vectors through feature extraction. Cross-rate, cross-pitch, and cross-frame features utilized in voice recognition applications were divided into groups using classifiers including decision trees, random forests, and k nearest neighbors. Linear prediction coefficients (LPC), Stabilized Audi-tory images (SAI), and Spectrogram image features (SIF) have all seen increased use in recent years. In a variety of applications, machine learning and soft computing methods including the Hid-den and Gaussian mixture model, random forests, multi-layer perceptron's, and forthcoming deep learning networks in sound are just a few examples. In recent years, SIF has been able to produce sound waves, allowing for more precise results in noisy environments. High- and low-pressure areas combine to form these sound waves as they travel through a medium. These alternating high- and low-pressure zones create a unique rhythm that allows us to identify individual sounds. The wavelength, frequency, wave speed, and interval of these waves are among their few distinguishing features [6]. Similar to how people do it, these features are used to categories sounds into distinct groups. The frequency range of a sound wave can be visualized using a spectrogram, as shown in Fig. 1. Simply put, it is a picture of the sound wave's frequency range [7].The sound signal's spectrogram is infre- quaintly produced, with the strongest components located at higher frequencies and the strongest noise at lower frequencies. Machine learning classifiers can be used in conjunction with the produced spectrogram images. Sun et al. [8] suggested a unified framework for Technique based on deep learning autoencoder and extreme machine learning models for spotting methods that use extreme learning to spot unstructured signal abstraction representations. Using a single-layer multi-perceptron model and non-linear activation functions, Baig et al. [9] suggested a method based on Ada-boosting. Unsupervised autoencoder deep networks are used in a number of cutting-edge applications, such as head and pose estimation, to detect a search-driven engine and its related non-linear mapping. The satellite pictures' semantic granular level representations were generated by a stacked discriminative sparse autoencoder**.**

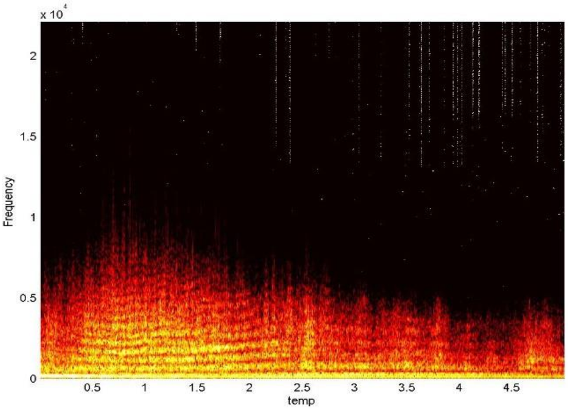

FIGURE 1. Generated spectrogram of a sound wave.

Liu et al. [10], in their comprehensive review of Deep learning models, found that the Convolutional Neural Network (CNN) worked best on image and video data. Even though CNNs are proficient at identifying objects in high-resolution remote sensing images [11], they struggle to do so when the objects have been rotated. To address this issue, a rotation-invariant CNN was suggested for object detection in sensory images. Features can be extracted and classified from spectrogram images using deep learning techniques because they provide a visual representation of the signal's frequency range. Less frequent sound signals have poor locality and produce varying spectrogram pattern representations. However, convolutional neural networks (CNNs) are becoming increasingly common in computer vision and audio processing due to their ability to classify spectrogram image features without being overly sensitive to the pattern's position on the generated spectrogram image. In the early 1990s, the first CNN structure was developed. When it comes to classifying handwritten numbers, LeNet-5 was the first Convolutional Neural Network to be created [12]. LeNet-5's results far surpassed those of competing methods at the time [13], [14]. CNN begins with a convolutional layer, which is responsible for attempting to understand the image's underlying features. The pooling layer follows, and its job is to lower the feature map's dimensions. The feature map is transferred from the convolutional layer to the pooling layer. Figure 2 illustrates how the number of convolutional layers and pooling layers can vary from one sample to the next. The ultimate convolutional laye

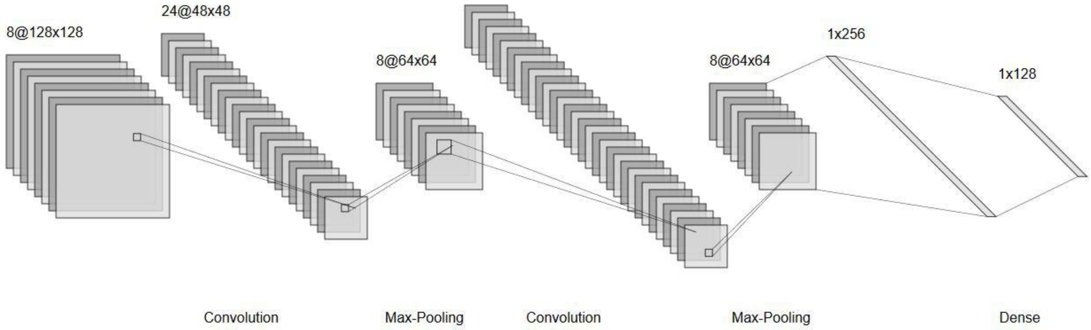

FIGURE 2. Convolutional neural network; as shown there are two layers of convolutional and pooling layer and a final dense layer.

II MATERIAL AND EXPERIMENT:

In this experiment, we use datasets that are very distinct from other available audio datasets. Here, instead of using voice datasets, we used environmental sound datasets. A major obstacle to creating an effective system for sound categorization is the scarcity of environmental sound datasets. For this experiment, we made use of the publicly accessible ESC-10 and ESC-50 datasets . There are a total of 400 environmental sound recordings across 10 groups included in the ESC-10 dataset. Dog barking, firecrackers, rain, rooster, infant cries, sneezing, sea waves, chainsaw, helicopter, and the ticking of a clock are just some of the sounds that fall into these categories.

A. EQUIPMENT AVAILABLE TO PERFORM THE EXPERIMENT

The Asus ROG Zephyrus GX501 notebook was used for the research. The following is a full list of this system's technical specifications: The computer has an Intel Core i7 processor, an Nvidia GeForce GTX 1080 graphics card with 8GB GDDR5X VRAM, and 16GB of RAM altogether.

B. APPLICATIONS USED

In this experiment, we used a convolutional neural network and a Tensor deep stacking network that we built and trained using a variety of tools, APIs, and libraries.

MATLAB: The primary goal of this project was to develop methods for producing spectrograms of noises included in datasets. The MATLAB function ‘spectrogram ()' was used to create a spectrogram of the audio input for this purpose. With a number of iterations in the loop equivalent to the number of samples in the dataset, themselves (gcf, 'name'. format)' function was used to save the generated spectrograms. Figure 4 displays the resulting spectrograms for the infant wailing class using this method.

1) ANACONDA It’s a free and open source Python distro with many useful machine learning add-ons. This programme makes setting up a virtual machine in Windows a breeze. The TWT Toolkit Tensor deep stacking network [30] is a free and open-source software package for building such a network. It comes with the majority of the necessary libraries to operate this toolkit. The tensor deep stacking network can be trained and evaluated with the help of this toolkit's many features. 2) KERAS This application programming interface [35] was developed to aid in the creation of deep neural networks. For this test, we built on top of TensorFlow and used the Kera’s library. The convolutional neural network used in the exercise was built using Kera’s. Keres’s arsenal of activations and optimizers makes incorporating them into the model a breeze. The system also made use of other libraries, such as NumPy and Scikit-Learn [36].

EXPERIMENT

1) A CONVOLUTIONAL NEURAL NETWORK The following specifications were utilized to create a sequence model in Keras, which was subsequently implemented using TensorFlow. As illustrated in Fig. 5, the convolutional neural network employed a two-layer deep architecture that included a fully connected output prediction layer. The code snippet representing the proposed implementation is shown in Figure 6. Figure 7 illustrates the comprehensive process of CNN processing. In the first convolutional layer, there were 32 filters of size 3x3 with ReLU activation [36]. The middle layer produced 32 feature maps. The complexity of the data was mitigated and filtered through a maximum pooling technique with a pooling size of 2x2.

III Discussions and Results:

The efficacy of the proposed approach is observed and evaluated with reference to various parameters, and the performance of the proposed CNN and TDSN based sound event classification system is compared to that of previously reported systems.

Cross-dataset-trained CNN. Table 1 compares the proposed CNN and TWT methods to prior studies using spectrogram image features for sound event classification. Table 1 shows that compared to the other systems, our method (CNN) gave 38.9%, 37.4%, 34%, 32.2%, and 29.76%, and TWT provided 21.9%, 22.4%, 17%, 15.2%, and 12.76% better performance. CNN Tanh and ReLU activation functions are used in the training process. The efficacy of the Tanh activation function is superior to that of the logistic sigmoid activation function, which it resembles. The Tanh function's primary benefit was mapping all the nearby zero inputs to the near-zero region. In contrast, it mapped all the negative inputs to a strongly negative region.

TABLE 1. Performance comparison of the proposed approach with other systems.

Spectrogram driven sound system	**Techniques/Architectures Used**	**Performance (Accuracy%)**
MFCC-SVM	Support Vector Machine (SVM)	34.1 %
MPEG-7	Decision Trees	33.6 %
Gabor	Random Forest	39.0 %
GTCC	K-Nearest Neighbors	40.8 %
MFCC-MP	Multi-Layer Perception	43.24 %
CNN	Convolutional Neural Network	73 %
CNN with Wavelet (Proposed)	CNN with Tensor Wavelet	85 %

Because it is half-rectified, ReLU activation function performed best in the experiment. ReLU has the disadvantage of mapping all negative inputs to zero. which cause a hindrance in training the network. The training and validation loss curve in CNN displays promising results with more data supplied to the training process. In contrast to training with 200 samples taken from the ESC-50 dataset, training loss was reduced to below 0.5 when 450 samples were used, as shown in Fig. 9. The capacity of CNN architecture to learn more features from huge datasets as opposed to small sample sizes is the cause of this performance.

Audio Signal: [<matplotlib.lines.Line2D at 0x7a0dc5556080>]

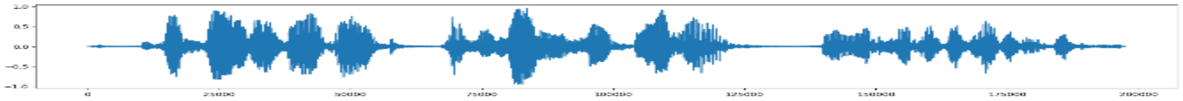

Figure 3:Audio Signal

ADD NOISE:[<matplotlib.lines.Line2D at 0x7a0dc3cabdc0>]

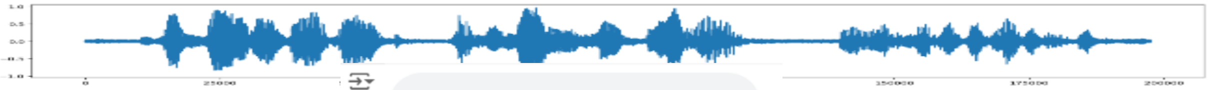

Figure 4:Noise Audio Signal

Reduce Noise:

Single band:

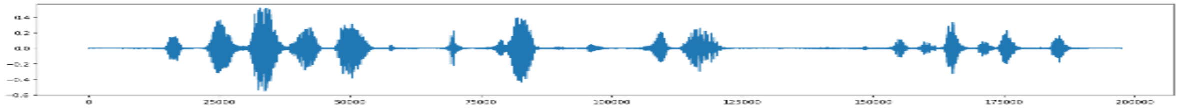

Figure 5: Reduced Noise Audio Signal

Double band:

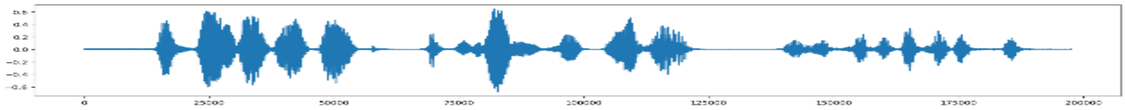

Figure 6: Reduced Noise Audio Signal

IV CONCLUSION:

The goal of this study was to evaluate how well the CNN and TDSN architectures performed when used to categories sound sources using spectrograms of the audible spectrum. Applications for convolutional neural networks that deal with picture categorization problems predominate. This work shows how sound classification using deep neural architectures is possible. This method for classifying sounds using spectrograms, CNN, and TWT required fewer trainable parameters than straight sound classification. We discovered that CNN and TWT had classification accuracy success rates of 85 % when comparing our trial results to those of other methods. It may be inferred from the experiment that this technique can offer trustworthy classification systems in the crucial regions. Whether CNN and Tensor Wavelet Network can accurately categories sound sources is an interesting research topic. By using the strength of tensors, the network may be trained on high resolution images rather than compressed images.

V REFERENCES :

References

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaten, ‘‘Content-based classifi- cation, search, and retrieval of audio,’’ *IEEE Multimedia*, vol. 3, no. 3,
pp. 27–36, Jun. 1996.

[2] F. Weninger and B. Schuller, ‘‘Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 337–340.

[3] M. V. Ghiurcau, C. Rusu, R. C. Bilcu, and J. Astola, ‘‘Audio based solutions for detecting intruders in wild areas,’’ *Signal Process.*, vol. 92, no. 3,
pp. 829–840, 2012.

[4] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, ‘‘Using one-class SVMs and wavelets for audio surveillance,’’ *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 763–775, Dec. 2008.

[5] S. Chu, S. Narayanan, and C.-C. J. Kuo, ‘‘Environmental sound recog- nition with time–frequency audio features,’’ *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[6] (2017). *Sound Classification.* [Online]. Available: http://www.paroc.com/knowhow/sound/sound-classification

[7] R. A. Altes, ‘‘Detection, estimation, and classification with spectrograms,’’
J. Acoust. Soc. Amer., vol. 67, no. 4, pp. 1232–1246, 1980.

[8] K. Sun, J. Zhang, C. Zhang, and J. Hu, ‘‘Generalized extreme learning machine autoencoder and a new deep neural network,’’ *Neurocomputing*, vol. 230, pp. 374–381, Mar. 2017.

[9] M. M. Baig, M. M. Awais, and E.-S. M. El-Alfy, ‘‘AdaBoost-based arti- ficial neural network learning,’’ *Neurocomputing*, vol. 248, pp. 120–126, Jul. 2017.

[10] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, ‘‘A survey of deep neural network architectures and their applications,’’ *Neurocomput- ing*, vol. 234, pp. 11–26, Apr. 2017.

[11] G. Cheng, P. Zhou, and J. Han, ‘‘Learning rotation-invariant convo- lutional neural networks for object detection in VHR optical remote sensing images,’’ *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12,
pp. 7405–7415, Dec. 2016.

[12] Y. LeCun, B. Boser, J. S. Denker, and D. Henderson, ‘‘Backpropagation applied to handwritten zip code recognition,’’ *Neural Comput.*, vol. 1, no. 4,
pp. 541–551, 1989.

[13] K. Simonyan and A. Zisserman, ‘‘Very deep convolutional networks for large-scale image recognition,’’ in *Proc. Int. Conf. Learn. Represent.*, 2015,
pp. 16–19.

[14] J. Gu *et al.*, ‘‘Recent advances in convolutional neural networks,’’ *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[15] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, ''Face recognition: A convolutional neural-network approach,'' *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[16] Y. LeCun and Y. Bengio, ''Convolutional networks for images, speech, and time series,'' *Handbook Brain Theory Neural Netw.*, vol. 3361, no. 10, p. 1995, 1995.

[17] D. Gupta and A. K. Ahlawat, ''Usability evaluation of live auction portal,'' *Cogn. Syst. Res.*, vol. 52, pp. 1036–1044, Dec. 2018.

[18] D. Gupta and A. K. Ahlawat, ''Usability feature selection via MBBAT: A novel approach,'' *J. Comput. Sci.*, vol. 23, pp. 195–203, Nov. 2017.

[19] Y. Bengio, ''Learning deep architectures for AI,'' *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[20] L. Deng, X. He, and J. Gao, ''Deep stacking networks for information retrieval,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3153–3157.

[21] T. G. Kolda and B. W. Bader, ''Tensor decompositions and applications,'' *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[22] B. Hutchinson, L. Deng, and D. Yu, ''Tensor deep stacking networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1944–1957, Aug. 2013.

[23] K. J. Piczak, ''Environmental sound classification with convolutional neu- ral networks,'' in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process.*, Sep. 2015, pp. 1–6.

[24] K. J. Piczak, ''ESC: Dataset for environmental sound classification,'' in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[25] D. Palzer and B. Hutchinson, ''The tensor deep stacking network toolkit,'' *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–5.

[26] (2018). *Keras Documentation*. [Online]. Available: https://keras.io/

[27] (2001). *SciPy: Open Source Scientific Tools for Python*. [Online]. Avail- able: https://www.scipy.org/

[28] A. L. Maas, A. Y. Hannun, and A. Y. Ng, ''Rectifier nonlinearities improve neural network acoustic models,'' in *Proc. Workshop Deep Learn. Audio, Speech, Lang. Process. (ICML)*, 2013 p. 3.

[29] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. (2012). ''Improving neural networks by preventing co-adaptation of feature detectors.'' [Online]. Available: https://arxiv.org/abs/1207.0580

[30] P. Tiwari *et al.*, ''Detection of subtype blood cells using deep learning,'' *Cogn. Syst. Res.*, vol. 52, pp. 1036–1044, Dec. 2018.

[31] D. Gupta and A. K. Ahlawat, ''Usability feature selection via MBBAT: A novel approach,'' *J. Comput. Sci.*, vol. 23, pp. 195–203, Nov. 2017.

[32] Y. Bengio, ''Learning deep architectures for AI,'' *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[33] L. Deng, X. He, and J. Gao, ''Deep stacking networks for information retrieval,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3153–3157.

[34] T. G. Kolda and B. W. Bader, ''Tensor decompositions and applications,'' *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[35] B. Hutchinson, L. Deng, and D. Yu, ''Tensor deep stacking networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1944–1957, Aug. 2013.

[36] K. J. Piczak, ''Environmental sound classification with convolutional neu- ral networks,'' in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process.*, Sep. 2015, pp. 1–6.

[37] K. J. Piczak, ''ESC: Dataset for environmental sound classification.