Audio IoT Analytics for Home Automation Safety

Sayed Khushal Shah
Computer Science Electrical Engineering
University of Missouri - Kansas City
Kansas City, USA
ssqn7@mail.umkc.edu

Zeenat Tariq
Computer Science Electrical Engineering
University of Missouri - Kansas City
Kansas City, USA
zt2gc@mail.umkc.edu

Yugyung Lee
Computer Science Electrical Engin.
University of Missouri - Kansas City
Kansas City, USA
leeyu@umkc.edu

***Abstract*—The aim of the paper is to perform audio analytics based on the audio sensor data that is continuously monitoring the home environment automatically through an audio Internet of Things (IoT) system. Domestic violence is one of the major problems in many cities nowadays. We have proposed a home automation system where IoT sensors records the audio in home environment continuously and the audio is sent to machine learning server where the audio is split into small clips and classified into different categories. The need of an automatic detection system is urgent for enforcing home safety and safe neighborhood. If IoT system detects any suspicious sound, it generates an emergency notification to nearest emergency services for possible action to be taken. The classification of audio such as gunshots, explosion, glass breaking, screaming and siren is based on shallow learning (Support vector machine, Decision tree, Random forest and Naïve Bayes) and deep learning (Convolutional neural network and Long short-term memory). Our experiments validated that Convolutional Neural Network shows the best performance (89% accuracy) compared to other machine learning algorithms.**

***Index Terms*—IoT Devices, Safe IoT Infrastructure, Deep learning, Audio Classification**

I. Introduction

During the last years, the smart city community show interests towards the concept of Artificial Intelligence and Internet of Things (IoT) in smart home and city context. Smart home and city are vision for the information and communication such as IoT that helps for improving the quality of living ensuring the economic and sustainable development of cities. Many wireless sensors have been deployed for the development of smart cities and IoT [1]–[3]. These sensors monitor the appropriate information related to the environment and simplify maintainable lifestyle. However, the large amount of data collected from these sensors need to be stored and processed properly to predict and classify the violence. Violence complaints are one of the major problems in most of the cities nowadays [4]. People suffer from violence issues in the cities, which affect the security system of neighborhood. Advanced technologies are required to analyze and tackle violence in smart cities. It can be done through the prediction of sound in smart home and city environments and make a system based on the classification of domestic violence data to provide a solution.

The demand for smart city applications for IoT is increased in past years. As time progresses, the interest of scientific community is going into the deeper level of making home and environment smart. The main purpose of smart home is to provide the technological access to city automation. This can make use of the mobile devices connected to the network to establish a smart home environment. IoT helps in making home smart by deploying sensors, which collect data in real time for the intelligent algorithm to take a bold decision on the available data to improve sustainability and quality of life.

Machine learning is one of the most promising techniques for the detection of domestic violence issues. Traditional machine learning that is shallow learning plays important role in classification of sounds. More recently, deep learning received attention due to their high performance in prediction and classification. Previous works been done are based on machine learning techniques to detect different type of violence in acoustic environment. These learning techniques are most growing fields nowadays in the area of audio classification [5]. DeepEar is audio sensing model for classification of audio for ambient scenes, emotion recognition, stress detection, and speaker identification in acoustic environment such as vehicles and caf [6]. Audio classification based on deep learning techniques [7] is related to environmental sound classification using Convolutional Neural Networks (CNN).

In this paper, we have proposed a solution, called smart domestic violence detection to tackle violence issues and make a successful smart home violence detection system. Our main contribution is divided into two main parts. In first part of the paper, we have created a model for audio mainly violence detection using IoT sensors. Based on this audio violence data, it has been observed that data analytics to predict types of domestic violence is required in home, where people are unaware of the surrounding happenings. We have put emphasis on violence data such as screaming, siren, explosion, gunshot and glass breaking that may happen when all neighbors are uninformed of the situation. The second part of our contribution is to design an automated detection system for domestic violence in such a way that if the users are away or even sleeping the system will have the capability to inform the police department for possible actions such as in situation of screaming. We have used machine learning approach to design a system to recognize the sound occurring and inform the police department when any suspicious sound happening around is detected. The focus of machine learning task is towards detection of domestic violence sounds such as gunshot, screaming, glass breaking, explosion and siren.

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

The rest of the paper is organized as follows. Section II describes the related work about machine learning models, IoT technology and domestic violence data. Section III explains the analysis of IoT sensor data, the design of the model, and the workflow of smart domestic violence detection system for smart home. Results and evaluation of our smart domestic violence detection model are described in Section IV. Section V explains the conclusion and future work.

II. Related Work

A. IoT Analytics

Sandulescu et al. [8] presented the idea to recognize stress level of person using sound features. They have designed an android application which has the capability to recognize the stress level by using a buffer recording session using shallow learning algorithm such as Support Vector Machine (SVM). The authors proposed to use microphone recording to improve the recognition of the stress level of a person. Stress can also cause someone to commit crime, however the author did not discuss that point in this paper. According our research we have more emphasis on IoT sensors rather than restricting to any platform. The sensors have the capability to detect sounds and recognize any suspicious activities related to violence. We have used machine learning algorithms for the classification of domestic violence related sounds.

Navarro et al. [9] mainly focused on environmental acoustics generated in cities. They considered noise as one of the major issues in town by collecting large data using the IoT sensors. However, the work emphasizes more on the collection of huge datasets. Our approach is based on designing of IoT system for home security, domestic violence data and the main attention is towards data analytics to predict and classify the domestic violence based on machine learning algorithms which are connected to a real-time detection system to inform police department automatically.

Zanella et al. [10] presented IoT enabled cities where they focused on urban area in the city of Padova, Italy where joint research was conducted with the city municipality. The survey on their IoT enabled services for communication technology, protocols, and architecture has paid the main attention towards the network architecture. Our approach is based on IoT setup and machine learning techniques, where we predict and classify the sound of domestic violence to take an appropriate action.

Nam et al. [11] considered the problem of smart city environment into three major things. First, they proposed the smart infrastructure for making a city smart. Second, they showed the involvement of people in social activities through social media platform. Third, they considered the governance of smart city project through institutions and the involvement of people in smart city project to facilitate the citizens. The authors discussed the concept of the interconnection between technology, human, and institutions in smart city. However, they did not consider any aspect of dealing with domestic violence issues and machine learning models for violence detection. Our focus is on domestic violence detection and classification in real time through machine learning algorithms.

B. Machine Learning Classification

Choi et al. [12] proposed Convolutional Recurrent Neural network (CRNN) for music classification. The last convolutional layers are replaced by Recurrent Neural Network (RNN), and both the classifiers are used for feature extraction and summarization, respectively. Computational controlled experiments were performed by changing the parameters of the networks.

Salamon and Bello presented the data augmentation technique [13] for environmental sound classification using Deep Convolutional Neural Network (DCNN). The deformation of audio was performed through time stretching, pitch shifting, dynamic range compression, and background noise.

Tang et al. [14] classified music genres using hierarchical Long Short Term Memory (LSTM) model. They have used a divide and conquer approach by dividing the ten genre musical data into a set of mild music and strong music dataset. The primary limitation was very small dataset and as the number of epochs increased overfitting occurred. They showed that Recurrent Neural Network (RNN) is more powerful network which reuses the parameters and learns from the previous states.

A multi label Recurrent Neural Network (RNN) has been proposed by Giambattista et al. [15] in the shape of bi directional Long Short Term Memory Recurrent Neural Network (LSTM-RNN) for polyphonic sound event detection in real life recordings. The dataset was augmented through time stretching, subframe time shifting and blocks mixing techniques and mapped to the Long Short Term Memory (LSTM) model. Piczak [7] proposed CNN for classification of environmental sounds. The architecture consists of two convolutional rectified layer unit by applying max pooling, two fully connected hidden layers, and a softmax output layer. The data was augmented through the random time delays and pitch shifting. Using *librosa* implementation, *Mel Spectrograms* were extracted from all audio files, resampled and normalized with different window sizes.

Saki et al. [16] highlighted the implementation of a system that captures the sounds in real time and is also tested on real field as an Android or iOS mobile application for classification of noise by applying Random Forest classifier. The emphasis was put on subband features. They covered the aspect of sound in signals, but they did not put any focus on other machine learning approaches apart from Random Forest. Our strategy in machine learning is mainly focused towards the comparison of several models in shallow learning and deep learning.

Atrey et al. [17] discussed the audio feature detection, when the traditional video surveillance fails due to dark. Audio can be considered as the vital feature at any time of the day. Previously, research has been done to detect cough sound or gunshot, but the authors claims that their research is better than the previous versions in the same field by considering human crying, shouting etc. They proposed

(a) Smart Home Environment

(b) IoT Audio Recognition System

Fig. 1. IoT Audio Analytics for Smart Home

a system which is using microphone sensors. Four features such as Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficient (LPCC) and Linear Frequency Cepstral Coefficient (LFCC) are extracted and GMM algorithm is applied for the classification of audio signals. By using GMM, they extracted the foreground and background sound and finally detecting excited and emotional sounds. They did comparison based on different audio features. While in our system we detect audio from IoT sensor and classify the sounds using machine learning techniques.

Although previous works were based on IoT, violence data and machine learning model, every model has specific limitation i.e. some of the work focus on classification while other worked on regression. Our approach is mainly focused on both shallow learning and deep learning models to perform a comparison with a real-time complaint system. We have used IoT sensors to detect sounds and finally considered domestic violence sounds to evaluate the proposed machine learning models. Based on the accuracy, a real-time domestic violence complaint will be submitted to the server.

III. IoT Audio Analytics for Smart Home

A. Smart Home Environment

Our model is based on real time solution to violence through sound detection using sensors and classification of domestic violence sounds using machine learning approach. The scope of the model does not depend on IoT solely. We performed real time testing using sensors and detection of sound. The domestic violence classification for smart home is based on machine learning. We are motivated to propose smart violence detection to provide home safety that will help in maintaining a safe neighborhood (as shown in Figure I(a)). The smart domestic violence detection system can detect the violence happening around and classifies the sound using machine learning algorithms.

B. Smart Home Audio Recognition Model

The architecture of smart domestic violence detection system is shown in Figure I(b). The smart domestic violence detection system has the capability to work in a smart home environment. During the design phase of the system, our focus was on the domestic violence and emergency issue that can be reported to the police department with minimum human intervention constrained by living in a neighborhood. For example, if a person is sleeping or unaware of any suspicious activity, the system has capability to recognize the violence sound and report it to the police department. Smart violence detection is capable enough to classify the sounds in the environment due to the machine learning model trained with the domestic violence data. The system intelligently detects the violence together with the intensity of the sound. An automatic complaint generated will be sent to the server with priority level indicating if a quick response to the incident is needed.

C. IoT Device Design

The smart domestic violence detection system is considered as a detection and complaint system for the resident when they are in the state of emergency. As shown in Figure 2, we have used Arduino pro mini version for our system design. The audio sounds are recorded in SD card with the system ability to split audio file into 10 second clips. We have used MAX9814 microphone which has the automatic gain control feature. MAX9814 amplifies the sound coming from a distance and controls very loud sound while considering noise cancellation. For demo purpose we have used LiPo battery which can give power back up of up to 8 hours to Arduino micro.

D. Machine Learning with the Audio Dataset

D.1. Workflow: Our classification for sound data was conducted based on the data analytics workflow with machine learning algorithms as shown in Figure 3. Machine learning algorithms play the main role in classifying sounds which falls

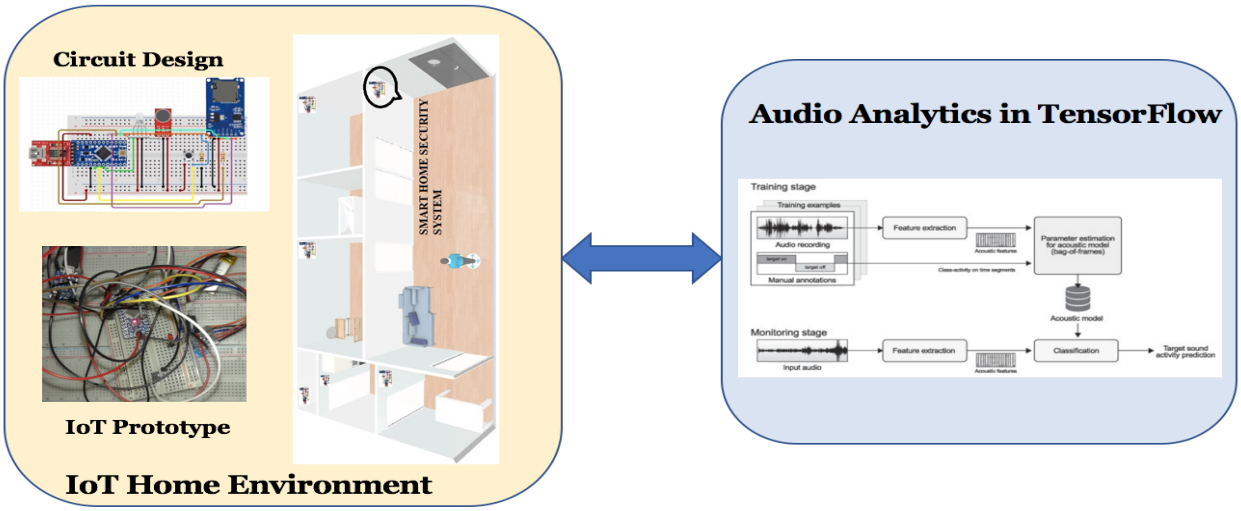

Fig. 2. IoT Device Prototype for Audio IoT Analytics

under the category of emergency and domestic violence such as *screaming*, *gunshots*, *siren*, *explosion*, and *glass breaking*. Machine learning component recognizes a type of violence based on the features extracted from the input audio data. Through domestic violence sound classification, complaint is sent to the server to respond according to the priority of the incident.

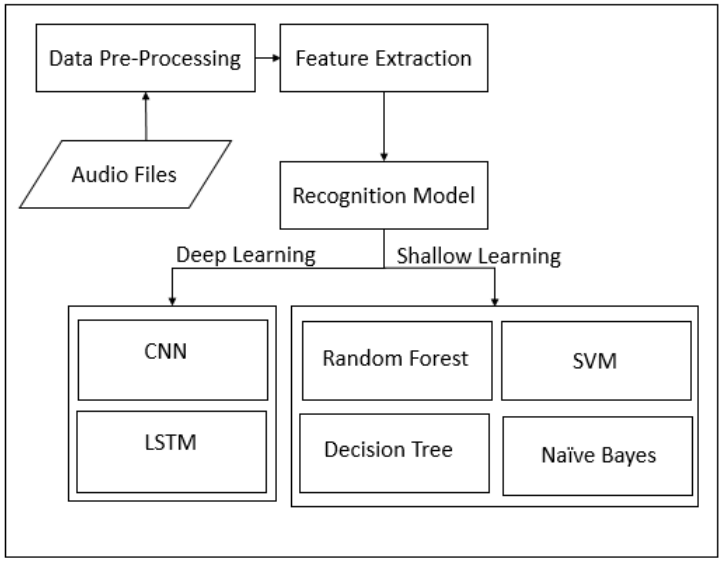

Fig. 3. Data Analytics Workflow

D.2. Feature Extraction: We have used librosa features [18]. The features extracted from the input audio data were *Mel Frequency Cepstral Coefficients (MFCC)*, *Chroma*, *Mel*, *Contrast*, and *Tonnetz*. Short Time Fourier Transform (STFT) is used to cut down the continuous signal into parts. Spectrograms operations such as inverse STFT and instantaneous frequency spectrogram are used for down streaming analysis of features [19]. *MFCC* is dominant for audio signal, speech recognition and is based on the Short Term Spectral Feature [20]. The Mel-scaled representation covers the musical tone, pitch and chroma that encode harmony by conquering loudness of signal. The *Chroma* implementations are delivered in two ways: fixed-window STFT and variable window constant-Q transform analysis. The *Tonnetz* function is another way of representing pitch and harmony, which is a geometric representation of pitch intervals derived from Chroma applied in audio and music [21].

D.3. Deep Learning: We have used two Deep Learning algorithms (CNN and LSTM) for classification of domestic violence sound. The use of different algorithms is required to find the best model, which can give us higher accuracy for the given dataset. The features are extracted from audio files separately, and these extracted features are processed by CNN and LSTM. Based on each algorithm the classification of domestic violence categories, for example, *gunshot*, *screaming*, *explosion*, *siren* and *glass breaking*, is performed. The performance comparison of the Deep Learning algorithms was compared with the Shallow Learning algorithms such as Random Forest, Support Vector Machine (SVM), Decision Tree, and Naïve Bayes.

Convolutional Neural Network (CNN) is a deep learning model applied on supervised learning comprised of convolutional layers with a subsampling step, followed by one or more layers as a fully connected layer. CNN has been an efficient and widely used model in the field of deep learning. The CNN based approach is used for classifying the classes rather than relying on manually engineered features for the recognition of the tasks, which include category classification, identifying the status of traffic signals or recognizing the exact house numbers [22].

The *Long short-term memory (LSTM)* network model can operate on recurrent memory blocks, in which one of the memory cells contains three multiplicative. The model can read, write, and reset the cell-based operation to avoid over-

fitting while learning by utilizing the temporal information of the cell for a specific duration of time.

Our analytical models are composed of two layers in the network, and we used our data to train a model with both layers. The first and second layers consist of 80 ReLU (Rectified Linear Unit) filters with max pooling on each layer and a stride size of 1 x 1. Finally, the training was performed using two fully connected hidden convolutional layers. We have used softmax as an activation for an output layer. For LSTM network, we considered two layers with Root Mean Square Propagation (RMSPROP) as an optimizer. In the first layer, we have used 128 neurons while the dropout is 0.05 and the recurrent dropout is 0.35. In the second layer, we considered 32 neurons with the same dropout and recurrent dropout. Finally, a fully connected layer with softmax as an activation function was applied.

IV. Result and Evaluation

A. Dataset

Based on our prototype, we have created a new dataset made out of free sounds [23] for classification. The dataset is based on violence categories, such as *gunshot*, *screaming*, *explosion*, *siren* and *glass breaking*. Each category has 150 audio data of 3 to 15 seconds each and has .wav format. For the audio dataset, we have set the batch size to 35. The data was split into 70% training and 30% testing with 1000 Epochs.

B. Feature Extraction Results

The model was designed in such a way that in first instance it extracts the five audio features from domestic violence audio data. The segmented information was used for analysis and the features extracted from librosa library [18] include *Short Term Fourier Transform*, *MFCC*, *Chroma*, *Tonnetz* and *Mel*. The extraction of MFCC features for each domestic violence category through the *librosa* library are shown in Figure 4, where the initially signal was turned into the frequency domain from the time domain through Fourier transformation. The power spectrum from frequencies were calculated and the Mel-filter banks were applied. After the filters, the Mel scale and logarithm functions were applied with the power spectrum and at the last the Cepstrum based on Discrete Cosine Transformation (DCT) is used to extract the change in frequencies.

C. Classification Results

In the evaluation, the proposed Deep Learning models (CNN and LSTM) were compared with the baseline shallow learning algorithms. The parameters of CNN and LSTM are shown in Table I. The baselines for the classification performance comparison include Random Forest, Support Vector Machine (SVM), Decision Tree, and Naïve Bayes.

Our deep learning models are implemented using Keras Tensorflow [24]. Table II shows the accuracy obtained from shallow learning and deep learning models. These models were tested for audio classification by dividing 30-70% testing and training. The best accuracy was obtained from Random Forest among the shallow learning algorithms (Random Forest,

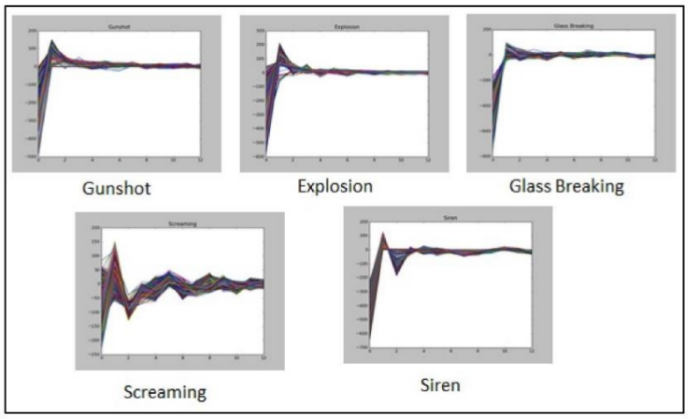

Fig. 4. MFCC Feature Extraction

TABLE I
Parameters for CNN and LSTM Models

Parameters	Value
Training:Testing	70:30
#Layers	2
#Epochs	1000
Total# Network Parameters	87,882
CNN dropout	0.05
LSTM recurrent dropout	0.35
Batch Size	35

Support Vector Machine, Decision Tree, and Naïve Bayes), which was 78%. The accuracy of each algorithm is shown in Figure 5. The maximum depth for Decision Tree and Random Forest classifiers was set to none that achieved the best accuracy. Using the dataset, we obtained the highest accuracy of 89% with CNN. LSTM obtained a lower accuracy than the CNN model since LSTM requires a large dataset for the network to be trained.

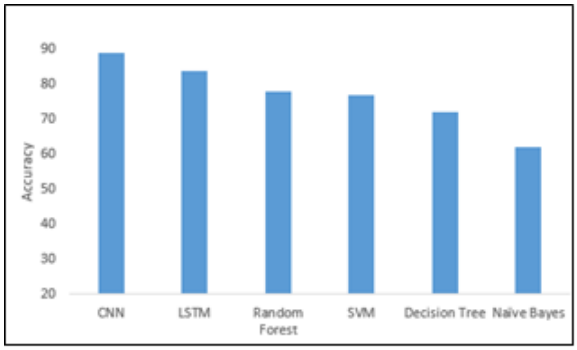

Fig. 5. Classification Accuracy

We have also trained our deep learning models on different numbers of epochs range from 100 to 1000. We found that the best accuracy with the CNN model was achieved when the number of epochs was reached 1000. For LSTM, the best accuracy was obtained when the number of epochs was reached above 800. The accuracy scores with the varying

TABLE II
MACHINE LEARNING PERFORMANCE

ML Type	Algorithm	Accuracy
Deep Learning	CNN	**89%**
	LSTM	85%
Shallow Learning	Random Forest	78%
	SVM	77%
	Decision Tree	72%
	Naïve Bayes	62%

epochs are shown in Figure 6. We have observed that when the number of epochs increases, the model performs overfitting and when we kept it low, it goes into underfitting.

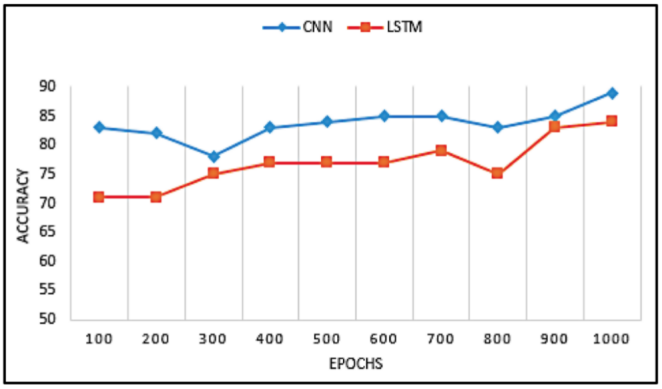

Fig. 6. Test Accuracy for Varying Number of Epochs

V. CONCLUSION AND FUTURE WORK

In this paper, we have used IoT sensors for detection of audio data for safe home environment. We developed the IoT system with the capability of sensing some types of domestic violence (screaming, gunshot, siren, glass breaking and explosion) and recognizing the sound using machine learning algorithms. We have also proposed a system that reports the incident to the police server for a possible action to be taken quickly. We performed audio classification with shallow learning (Random Forest, Support Vector Machine, Decision Tree, and Naïve Bayes) and deep learning (CNN and LSTM). We have observed that CNN provides the best accuracy for audio detection.

The smart home automation system can be integrated with the services of the police and fire departments for explosion. The smart sensing with camera sensors can be used to detect and report the specific incidents with sounds or footage to the concerned departments using video analytics. We will have more data available for an intelligent system to be trained. The state departments can have a standard format for data to make future research possible. The IoT sensors can be connected to the local or remote server for more detailed analysis. In this way, we believe that we can make homes and neighborhood more safe and secure.

REFERENCES

[1] K. Su, J. Li, and H. Fu, "Smart city and the applications," in *Electronics, Communications and Control (ICECC), 2011 International Conference on*. IEEE, 2011, pp. 1028–1031.

[2] M. Conner, "Sensors empower the" internet of things"," *EDN (Electrical Design News)*, vol. 55, no. 10, p. 32, 2010.

[3] I. F. Akyildiz and M. C. Vuran, *Wireless sensor networks*. John Wiley & Sons, 2010, vol. 4.

[4] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and vision computing*, vol. 48, pp. 37–41, 2016.

[5] L. Deng, D. Yu *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[6] N. D. Lane, P. Georgiev, and L. Qendro, "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 283–294.

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

[8] V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, and O. M. Mozos, "Stress detection using wearable physiological sensors," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2015, pp. 526–532.

[9] J. Navarro, J. TomasGabarron, and J. Escolano, "On the application of big data techniques to noise monitoring of smart cities."

[10] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things journal*, vol. 1, no. 1, pp. 22–32, 2014.

[11] T. Nam and T. A. Pardo, "Conceptualizing smart city with dimensions of technology, people, and institutions," in *Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times*. ACM, 2011, pp. 282–291.

[12] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," *arXiv preprint arXiv:1609.04243*, 2016.

[13] 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, 2017.

[14] C. P. Tang, K. L. Chui, Y. K. Yu, Z. Zeng, K. H. Wong *et al.*, "Music genre classification using a hierarchical long short term memory (lstm) model," 2018.

[15] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," *arXiv preprint arXiv:1604.00861*, 2016.

[16] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz, "Smartphone-based real-time classification of noise signals using subband features and random forest classifier," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2204–2208.

[17] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.

[18] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[19] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 756–759.

[20] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *ISMIR*, vol. 270, 2000, pp. 1–11.

[21] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Moving beyond feature design: Deep architectures and automatic feature learning in music informatics." in *ISMIR*. Citeseer, 2012, pp. 403–408.

[22] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3288–3291.

[23] Alastair. Freesound dataset - explore. [Online]. Available: https://datasets.freesound.org/fsd/explore/

[24] N. Ketkar, "Introduction to keras," in *Deep Learning with Python*. Springer, 2017, pp. 97–111.