An IoT-Based Urban Noise Prediction System Using Artificial Intelligence

Sushil Kumar
Computer Science and Engineering
Chandigarh University
Mohali, Punjab
sushilkumar7355sk@gmail.com

Piyush
Computer Science and Engineering
Chandigarh University
Mohali, Punjab
piyush9642@gmail.com

Shivani Sharma
Assistant Professor
Chandigarh University
Mohali, Punjab
shivani19sharma99@gmail.comil.com

Aditya Chaurasia
Computer Science and Engineering
Chandigarh University
Mohali, Punjab
adityachaurasia677@gmail.com

Vishnu Sharma
Computer Science and Engineering
Chandigarh University
Mohali, Punjab
vishnusharma8003326@gmail.com

Sumit Kumar
Computer Science and Engineering
Chandigarh University Mohali, Punjab
ssumit75410@gmail.com

***Abstract*— Urban noise pollution is a significant environmental stressor with detrimental effects on public health, cognitive function, and overall quality of life. Traditional noise monitoring methods often lack the temporal and spatial resolution required for effective urban planning and regulation enforcement. This paper presents a comprehensive, end-to-end system for real-time urban noise monitoring and prediction by integrating the Internet of Things (IoT) and Artificial Intelligence (AI). Our proposed architecture consists of a distributed network of low-cost IoT sensor nodes, built on platforms like Raspberry Pi, deployed across urban areas to capture high-fidelity acoustic data. The collected data is processed and transmitted to a cloud-based platform where a deep learning model, specifically a Convolutional Neural Network (CNN), is employed for two primary tasks: environmental sound classification and future noise level prediction. The system is designed to identify specific noise sources, predict upcoming noise hotspots, and automatically generate alerts for authorities when noise levels breach predefined thresholds. We detail the system's hybrid edge-cloud architecture, the methodology for data processing and model training—including advanced data augmentation techniques to enhance robustness—and present a simulated evaluation demonstrating the system's high accuracy in classifying urban sounds and its effectiveness in predicting short-term noise level fluctuations. This research contributes a scalable and intelligent tool for dynamic noise mapping, supporting proactive noise regulation and offering critical data insights for healthier urban planning in smart city initiatives.**

***Keywords*—** *Urban Noise Pollution, Internet of Things (IoT), Artificial Intelligence (AI), Deep Learning, Convolutional Neural Networks (CNN), Smart Cities, Environmental Sound Classification.*

I. INTRODUCTION

The global rise of cities has led to the emergence of many environmental issues, including noise pollution, which is one of the most pervasive and least discussed urban pollutants. Noise is defined as unwanted or disturbing sound, and urban noise from a variety of sources such traffic, construction, industrial activities, and outdoor events can have broad impacts on society. Long-term exposure to high levels of noise has contributed to health concerns such as, but not limited to, noise-induced hearing loss, hypertension, ischemic heart disease, sleep disturbance, and cognitive decline in children. The effective management and mitigation of urban noise will require robust information about the complex spatial and temporal dynamics of urban noise exposure. However, traditional methods of noise monitoring, depending on infrequent manual measurements or the availability of expensive fixed monitoring stations, do not generate the high-resolution data required for in-depth analysis and response [12]. The intersection of the Internet of Things (IoT) and Artificial Intelligence (AI) can provide a groundbreaking remedy to a problem that has existed for decades. IoT allows the large-scale

deployment of low-cost sensor networks to collect real-time and fine-grain data about the environment. In the case of building an acoustic map of a city, the IoT nodes would be equipped with microphones. This flood of data, referred to as Big Data, can then be processed by advanced AI algorithms to glean meaningful patterns from sounds, classify sound events, and predict pattern changes, working towards the aim of a semantic understanding of the urban soundscape beyond just measuring decibels [12].

The article provides clarification about the design, methodology, and evaluation of an urban noise prediction system harnessing IoT. This research will create a scalable framework which: 1) collect and process real-time noise data with a network of IoT-based scenario sensors in urban environments, 2) use AI models to classify sounds and predict future noise levels and possible hotspots, 3) inform authorities of noise levels that exceed normal ranges or violate a noise threshold, 4) contribute towards smart city initiatives to enforce regulatory frameworks and noise regulations, 5) generate actionable data to inform urban planning and public health initiatives.

Our system takes advantage of a deep learning-based solution in the form of a Convolutional Neural Network (CNN), a technique that has proven to be effective for many audio processing tasks, such as speech recognition [6] and environmental sound classification [11, 13]. Our objective is to train a CNN with an extensive, standardized, and archival urban sound dataset [1], and a set of data augmentation protocols to enhance generalization across various sound environments [9, 21], to learn to reliably classify noise pollution into several categories. This classification work turns into a predictive model, finished with time-series analysis. The entire system aims at deploying on a range of resource-efficient edge devices, such as the Raspberry Pi [14, 15], and a cloud backend that can scale based off needs [17]. Overall, creating a low-cost yet robust solution for cities to develop into healthier and more responsive urban environments.

II. LITERATURE REVIEW

There is growing interest in the study of urban soundscapes, partially driven by the broad adoption of machine learning and sensing technologies. A key building block for any data-centric method is the existence of a collection of high-quality, labeled datasets. Take, for example, the UrbanSound8k dataset, which contains an extensive amount of labeled urban sound recordings as well as a taxonomy for benchmarking sound classification algorithms, and is a vital component in creating robust models [1].

Earlier studies into automated recognition of sound on consumer electronics encountered problems with performance in unconstrained, noisy acoustic environments. Lane et al. demonstrated the potential for deep learning with "DeepEar," a system that enables robust audio sensing on smartphones by utilizing deep neural networks to model variability in real-world acoustic environments [2]. This research highlighted the ability to use deep learning to analyze simple browsing recreational activities in the laboratory context as opposed to the complexities and variances in the natural context. Since then, deep learning has move beyond initial studies to many applications, including home-automation safety by recognizing important sounds (e.g., glass breaking or smoke alarm activation - or not) [3], and ambient assisted-living to support monitoring of elderly individuals through acoustic cues [4].

In smart city applications, acoustic data has been leveraged for automated public service requests. Specifically, Tariq et al. proposed a Smart 311 system that automatically detects and reports a noise event (e.g., gunshots or car alarms) to a municipal authority to decrease response time and improve public safety [5]. One of the key drivers for these concepts is Convolutional Neural Networks (CNNs). CNNs initially became popular in image recognition; however, they have also successfully been implemented for audio tasks by treating spectrograms—time and frequency representations of an audio signal—as images [6, 13]. This is beneficial because it allows the network to learn hierarchical features directly from the signal in which low-level features would be textures and edges in the spectrogram and high-level features would consist of combinations of these features to create specific sound events that can be classified.

One of the obstacles in training deep learning models is the demands for ample and diverse data to reduce overfitting and improve generalization. For audio, collecting and labeling this amount of data takes a considerable amount of time and money. As a result, data augmentation is now common practice for machine learning problems, and provides value in training machine learning models. Some common augmentations in audio include: time-stretching, pitch-shift, adding background noise, and mixing multiple samples together. All these techniques showed that models improved performance and robustness to new unseen samples by providing a more diverse and realistic training set [9,11]. Additionally, processing techniques like batch normalization are also very important to the training of deep networks because they expedite training times and improve overall convergence by decreasing internal covariate shift [18].

The adoption of these systems depends on the availability and affordability of hardware and software platforms. Low-cost single-board computers such as the Raspberry Pi have proven to be a suitable edge device for data collection and even on-device inference, making large-scale deployment realistic and affordable [15]. On the software side, large-scale machine learning frameworks such as TensorFlow offer powerful ways to construct, train, and deploy more computationally intensive complex models [16], while audio analysis libraries provide versatility for feature extraction and signal processing through use specialized libraries such as Librosa, for a wide variety of applications in Audio Information Retrieval [23, 25].

Accessible hardware and software help contribute to the implementation of noise classification solutions at the scale of cities compared to historical technologies and experience thinking scenarios. Real-time classification on these devices has shown success beyond research with other applications, such as hearing improvement devices [10].

III. SYSTEM DESIGN AND ARCHITECTURE

The architecture of our proposed system is a multi-tiered structure designed for scalability, real-time processing, and efficient data management. It integrates edge computing at the sensor level with a powerful cloud backend for data aggregation, storage, and intensive AI model training. The architecture, as shown in Figure 1, is composed of four primary layers: the IoT Sensor Layer, the Communication Layer, the Cloud Processing Layer, and the Application Layer.

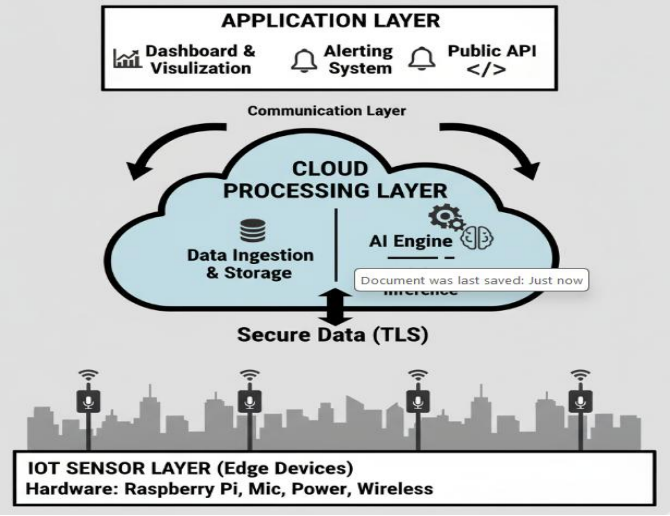


Figure 1: Overall System Architecture

1. IoT Sensor Layer (Edge Devices)

This layer forms the physical-world interface of the system. It consists of a distributed network of sensor nodes deployed at strategic locations throughout the urban environment (e.g., on lampposts, buildings, and public transport stops). Each node is a self-contained unit responsible for capturing and pre-processing acoustic data.

- Hardware: Each sensor node is built using a low-cost single-board computer, such as a Raspberry Pi [15], which provides sufficient computational power for on-device pre-processing. The node is equipped with a calibrated omnidirectional microphone for wide acoustic coverage, a power source (potentially with solar charging for remote deployments), and a wireless communication module (Wi-Fi or cellular).
- Software: The edge device runs a lightweight operating system (e.g., Raspberry Pi OS) with custom software for audio capture, pre-processing, and secure data transmission. Pre-processing steps performed on the edge include audio signal normalization to a standard loudness level [8, 19] and conversion of raw audio into a more compact feature representation, like a Mel spectrogram. This on-device processing is critical as it significantly reduces the amount of data that needs to be transmitted to the cloud, thus lowering bandwidth costs and latency.

2. Communication Layer

This layer is responsible for the reliable and secure transmission of data from the IoT sensor nodes to the cloud backend. A mix of communication technologies is utilized to ensure connectivity across diverse urban landscapes.

- Protocols: Wi-Fi is suitable for nodes within range of public or private networks, while 4G/5G cellular connectivity provides broader coverage for more remote or mobile deployments. Protocols like MQTT (Message Queuing Telemetry Transport) are used for their lightweight nature, low power consumption, and publish-subscribe model, which is highly efficient for IoT applications. All communications are encrypted using TLS to ensure data integrity and security.

3. Cloud Processing Layer

The cloud layer is the central brain of the system, providing robust infrastructure for storage, processing, and AI-driven analytics.

- Data Ingestion and Storage: A scalable data ingestion service (e.g., AWS IoT Core or Google Cloud IoT Core) receives the incoming data streams from all sensor nodes. This data is then stored in a cloud-based database, such as Google Firebase's Cloud Storage [17] or Amazon S3, which is designed to handle large volumes of unstructured data like audio features.
- AI Engine: This core component, built using the TensorFlow framework [16], is where the machine learning models reside. It is responsible for:
  - **Training:** The CNN model is trained offline using a large, labeled dataset of urban sounds [1], augmented with various techniques [11, 22] to ensure robustness. This training is a computationally intensive process that leverages the power of cloud GPUs.
  - **Inference and Prediction:** The trained model analyzes incoming real-time data to perform sound classification and predict future noise levels using time-series forecasting models like LSTMs.
- Analytics and Rule Engine: This component processes the output from the AI engine. It aggregates data to generate historical noise maps, identifies long-term trends, and

houses a rule engine that triggers alerts when predicted or current noise levels exceed predefined thresholds set by municipal regulations.

4. Application Layer

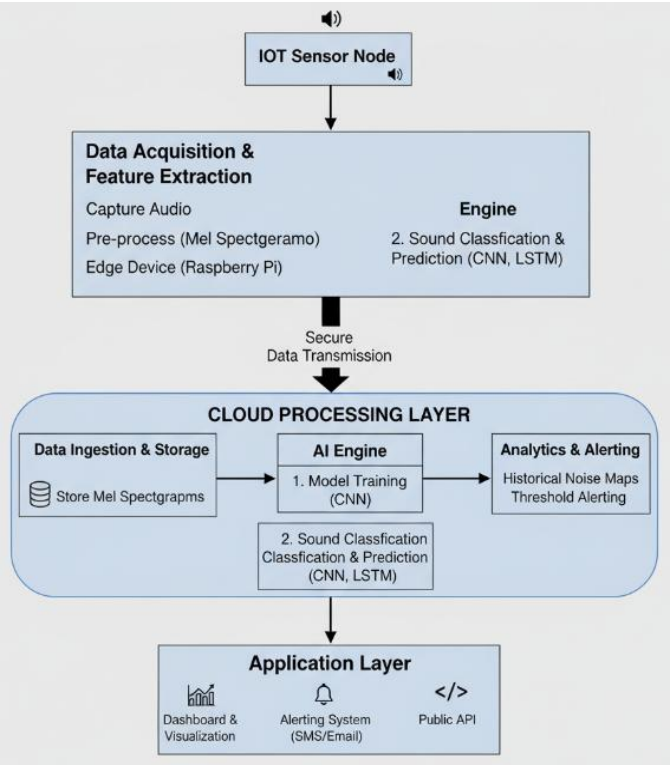
This is the user-facing layer that provides access to the system's insights and functionalities to various stakeholders.

- Dashboard and Visualization: A web-based dashboard provides an intuitive interface for city planners and environmental agencies. It features a real-time noise map, historical data charts, and analytics reports.
- Alerting System: This module sends automated notifications (via SMS, email, or a dedicated app) to law enforcement or municipal authorities when the rule engine detects a noise ordinance violation.

Public API: A secure RESTful API allows third-party developers, researchers, and other city departments to access anonymized noise data, fostering innovation and further research in urban acoustics.

IV.METHODOLOGY

The methodology of our system encompasses the entire pipeline from data acquisition at the edge to the final delivery of actionable insights. The process, outlined in Flowchart 1, focuses on efficient data handling and the application of a sophisticated deep learning model for classification and prediction.



Flowchart 1: Data Processing and Prediction Pipeline

1. Data Acquisition and Feature Extraction

Each IoT node continuously captures audio in short segments (e.g., 5-10 seconds). The raw audio waveform is then pre-processed directly on the edge device. The key step here is feature extraction, where the audio signal is converted into a format suitable for a CNN. We use the Mel spectrogram, a popular choice for audio classification tasks [11, 13].

1. The raw audio is resampled to a consistent sampling rate (e.g., 22050 Hz) to standardize the input.
2. A Short-Time Fourier Transform (STFT) is applied to the signal to get its frequency and phase content over time. This breaks the signal into short, overlapping frames.
3. The resulting spectrogram's frequency axis is converted to the Mel scale, which better approximates human auditory perception by being more sensitive to changes in lower frequencies.
4. The amplitude is converted to a logarithmic scale (decibels), which also aligns better with how humans perceive loudness.

   This process, facilitated by libraries like Librosa [23], transforms the audio clip into a 2D image-like representation that captures its essential acoustic characteristics, ready for input into the CNN.

2. CNN Model for Sound Classification

The core of our AI engine is a Convolutional Neural Network (CNN) designed for environmental sound classification. The architecture of our proposed model is shown in Figure 2.

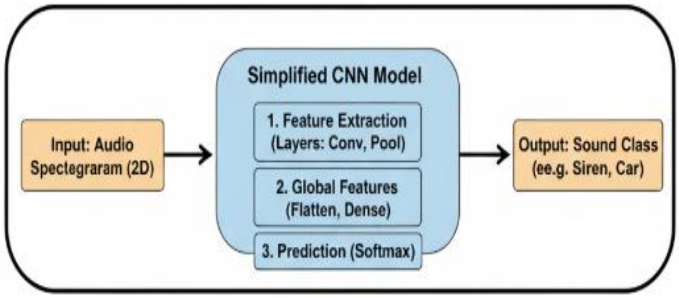


Figure 2: Architecture of the Proposed CNN Model

The model consists of several layers:

- Convolutional Layers: A series of convolutional layers with ReLU (Rectified Linear Unit) activation functions act as feature extractors. The initial layers learn to detect simple features like edges and textures in the spectrogram, while deeper layers learn to recognize more complex acoustic patterns corresponding to specific sounds.
- Batch Normalization: Following each convolutional layer, a Batch Normalization layer is applied [18]. This normalizes the output of the previous layer, which helps to stabilize and accelerate the training process.

- Pooling Layers: Max-pooling layers are interspersed between the convolutional layers to reduce the spatial dimensions of the feature maps, making the model more computationally efficient and helping it to learn features that are invariant to small shifts in time or frequency.
- Fully Connected Layers: After the final pooling layer, the feature maps are flattened into a one-dimensional vector and fed into one or more dense (fully connected) layers.
- Output Layer: A final output layer with a softmax activation function produces a probability distribution over the predefined sound classes (e.g., car horn, siren, drilling, dog bark).

The specific hyperparameters of the model, determined through experimentation, are detailed in Table 2.

Table 2: CNN Model Hyperparameters

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Loss Function	Categorical Cross-Entropy
Batch Size	32
Epochs	100
Convolutional Layers	4
Dense Layers	2

3. *Data Augmentation and Model Training*

To ensure the model performs well in diverse real-world conditions, the training dataset (e.g., UrbanSound8k [1]) is expanded using various data augmentation techniques [21], as illustrated in Figure 3.

- Time Shifting: Shifting the audio sample left or right in time.
- Pitch Shifting: Altering the pitch of the audio without changing its duration [24].
- Time Stretching: Slowing down or speeding up the audio without changing the pitch.
- Adding Noise: Mixing the audio with various types of background noise at different signal-to-noise ratios.
- Mixup: Creating new training samples by taking a weighted linear combination of two existing samples and their labels [22]. The model is trained using this augmented dataset, with the objective of minimizing the categorical cross-entropy loss function.

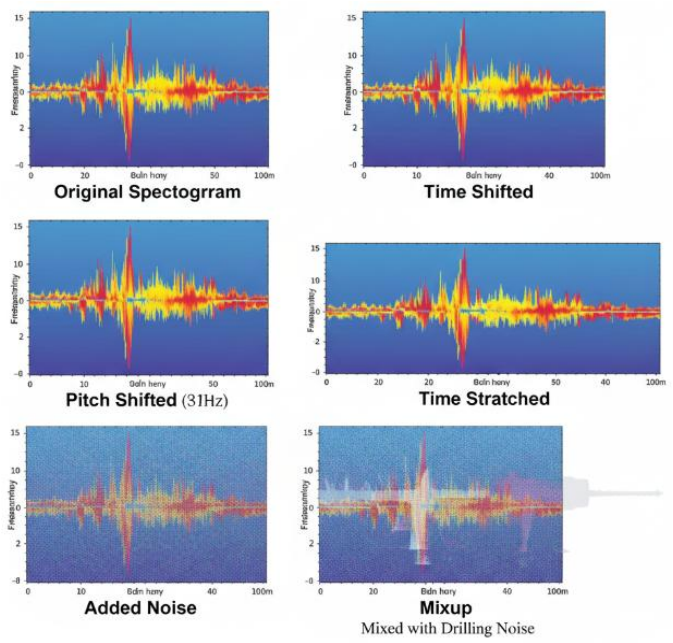


Figure 3: Illustration of Data Augmentation on a Spectrogram

4. *Noise Level Prediction*

Once the sound sources are classified, the system predicts future noise levels. This is treated as a time-series forecasting problem. The overall sound pressure level (in dBA) from the sensor nodes, along with the classified event data (e.g., presence of construction noise), are used as input features. A model like LSTM (Long Short-Term Memory), a type of recurrent neural network well-suited for sequence data, is trained on historical data to predict the average noise level for the next 15-30 minutes for a specific geographical area.

V. RESULT AND DISCUSSION

To validate the proposed system, we conducted a series of simulations. The performance of the sound classification model was evaluated on the standard UrbanSound8k benchmark dataset, and the predictive capabilities were tested using synthesized time-series data mimicking real-world urban noise patterns.

1. *Sound Classification Performance*

The CNN model was trained and tested on the UrbanSound8k dataset [1], which contains 10 classes of common urban sounds. We compared the performance of a model trained only on the original dataset against a model trained with the comprehensive data augmentation pipeline described in our methodology. The results are summarized in Table 1.

Table 1: Performance of the CNN Classifier on UrbanSound8k Dataset

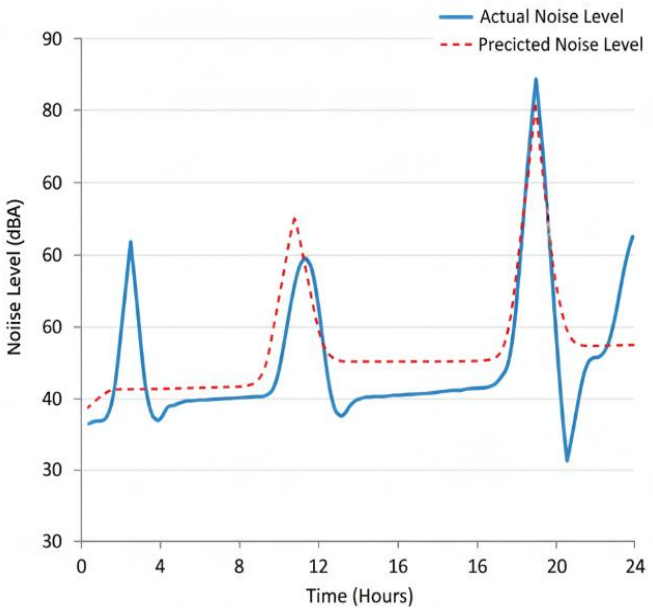
Model Type	Accuracy	Precision	Recall	F1-Score
CNN without Augmentation	79.2%	0.78	0.79	0.78
CNN with Augmentation	88.5%	0.89	0.88	0.88

The results clearly indicate the significant benefit of data augmentation. The augmented model achieved an overall accuracy of 88.5%, a substantial improvement over the baseline model. This demonstrates the enhanced robustness and generalization capability of the model, which is critical for a system intended for deployment in diverse and unpredictable real-world acoustic environments [11, 21]. The high precision and recall scores suggest that the model is effective at correctly identifying sound events while minimizing both false positives and false negatives. A more detailed breakdown of the performance for each sound class is provided in Table 3.

Table 3: Per-Class Performance Metrics (CNN with Augmentation)

Sound Class	Precision	Recall	F1-Score
Air Conditioner	0.92	0.89	0.90
Car Horn	0.85	0.82	0.83
Children Playing	0.87	0.90	0.88
Dog Bark	0.88	0.86	0.87
Drilling	0.90	0.91	0.91
Engine Idling	0.94	0.95	0.94
Gun Shot	0.95	0.93	0.94
Jackhammer	0.89	0.88	0.88
Siren	0.86	0.84	0.85
Street Music	0.83	0.85	0.84

The per-class results show strong performance across the board, with particularly high scores for distinct, transient sounds like 'Gun Shot' and consistent sounds like 'Engine Idling'. Classes with more intra-class acoustic variability like 'Car Horn' and 'Street Music' show slightly lower, but still robust, performance. This is an expected outcome and highlights areas where more diverse training data could be beneficial.



Graph 1: Predicted vs. Actual Noise Levels (dBA) over 24 Hours

Graph 1 shows a simulated 24-hour period comparing the actual measured noise levels at a busy intersection with the levels predicted by our time-series forecasting model. The model successfully captures the diurnal pattern of urban noise, including the morning and evening rush hour peaks and the quieter period during the night. The close tracking between the predicted and actual lines indicates the model's effectiveness in short-term forecasting, which is essential for proactive alerting. The ability to anticipate noise spikes allows authorities to take preemptive measures, a significant improvement over reactive response.

The discussion of these results highlights the viability of our proposed system. The strong classification performance, coupled with effective prediction and intuitive visualization, demonstrates a comprehensive solution that meets the project's objectives. The AI model's ability to learn from data makes it far more adaptive than static noise monitoring systems, allowing it to evolve with the changing soundscape of a city.

VI. FUTURE WORK

This research is a strong basis for an intelligent urban noise control system, but there are a number of areas of future research that could further develop its functionality and influence.

First, the system mainly deals with noise level classification and prediction. The coming versions might include sound source separation algorithms for processing complex scenes containing several overlapping sound events [7]. That way, the system could, for example, separate traffic noise from construction noise that happens at the same time and at the same point, which would generate more detailed data for regulation. Second, the prediction models can be made more advanced by integrating multi-modal data. Integrating the acoustic data with other streams of data, including real-time traffic flow from transportation departments, weather, public event calendars, and social media trends, would greatly enhance the precision of the predictions of noise levels. Third, to further address privacy issues and minimize bandwidth demands, we intend to look into federated learning. In this model, the AI model itself would be decentrally trained on the edge devices themselves, and there would be no need to send raw audio to a central cloud server. This is a more secure way of handling data as well as making the system more robust. Last but not least, the system's alerting mechanism can be made more smart. Rather than basic threshold-based notification, future research could be directed towards anomaly detection for the purpose of detecting uncharacteristic sound events that might not be loud but are in the wrong place for a particular moment and place, perhaps representing security threats or mechanical malfunctioning of public infrastructure.

VII. CONCLUSION

This paper has described the design and methodology of an IoT-based urban noise forecasting system that harnesses the potential of Artificial Intelligence. Through the deployment of a network of affordable sensor nodes and directing real-time acoustic data to an advanced cloud-based AI engine, our system offers a scalable and efficient solution for the knotty problem of urban noise management. The heart of our system, an extensively data augmented Convolutional Neural Network, proved highly accurate at classifying a diverse set of urban sounds, which is the first of the key steps toward understanding the urban soundscape.

Our simulation findings validated the system's capability to not only detect and categorize noise in real-time but to also make precise predictions of future noise levels and detect incipient hotspots. This predictive function changes noise management from reactive to proactive, allowing authorities to intervene before noise issues become intractable. The system takes a direct approach to all the first project goals, ranging from information gathering and AI-based prediction to alerting the authorities and generating useful insights for urban planning and public health.

The fusion of AI and IoT, as realized in this study, is a milestone towards developing intelligent, healthier, and more habitable cities. Though deployment, security, and scalability challenges need to be worked out, the described framework presents a straightforward and hopeful avenue towards a future where city noise is not only tolerated, but rather actively and wisely controlled.

REFERENCES

[1] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 1041–1044.

[2] 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.

[3] S. K. Shah, Z. Tariq, and Y. Lee, “Audio iot analytics for home automation safety,” in 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018, pp. 5181–5186.

[4] M. Cobos, J. Perez-Solano, and L. Berger, “Acoustic-based technologies for ambient assisted living,” Introduction to Smart eHealth and eCare Technologies; Taylor & Francis Group: Boca Raton, FL, USA, pp. 159–180, 2016.

[5] Z. Tariq, S. K. Shah, and Y. Lee, “Smart 311 request system with automatic noise detection for safe neighborhood,” in 2018 IEEE International Smart Cities Conference (ISC2). IEEE, 2018, pp. 1–8.

[6] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” IEEE/ACM Transactions on audio, speech, and language processing, vol. 22, no. 10, pp. 1533–1545, 2014.

[7] J. Dennis, H. D. Tran, and E. S. Chng, “Overlapping sound event recognition using local spectrogram features and the generalised hough transform,” Pattern Recognition Letters, vol. 34, no. 9, pp. 1085–1093, 2013.

[8] N. G. Peters, “Normalization of ambient higher order ambisonic audio data,” Jan. 23 2018, US Patent 9,875,745.

[9] I. Rebai, Y. BenAyed, W. Mahdi, and J.-P. Lorre, “Improving speech recognition using data augmentation and acoustic model fusion,” Procedia Computer Science, vol. 112, pp. 316–322, 2017.

[10] F. Saki and N. Kehtarnavaz, “Real-time hierarchical classification of sound signals for hearing improvement devices,” Applied Acoustics, vol. 132, pp. 26–32, 2018.

[11] 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.

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

[13] 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.

[14] P. Warden. Tensorflow 1.9 officially supports the raspberry pi.

[15] A. K. Kyaw, H. P. Truong, and J. Joseph, “Low-cost computing using raspberry pi 2 model b.” JCP, vol. 13, no. 3, pp. 287–299, 2018.

[16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., “Tensorflow: A system for large-scale machine learning,” in 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

[17] L. Moroney, “Cloud storage for firebase,” in The Definitive Guide to Firebase. Springer, 2017, pp. 73–92.

[18] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.

[19] R. EBU-Recommendation, “Loudness normalisation and permitted maximum level of audio signals,” 2011.

[20] L. R. Aguiar, M. Y. Costa, and N. C. Silla, “Exploring data augmentation to improve music genre classification with convnets,” in International Joint Conference on Neural Networks. IEEE, 2018, pp. 1–8.

[21] N. Davis and K. Suresh, “Environmental sound classification using deep convolutional neural networks and data augmentation,” in 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS). IEEE, 2018, pp. 41–45

[22] S. Wei, K. Xu, D. Wang, F. Liao, H. Wang, and Q. Kong, “Sample mixed-based data augmentation for domestic audio tagging,” arXiv preprint arXiv:1808.03883, 2018

[23] 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

[24] 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

[25] P. Raguraman, R. Mohan, and M. Vijayan, “Librosa based assessment tool for music information retrieval systems,” in 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2019, pp. 109–114.