PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0315

Soundscapes on edge - The real-time machine learning approach for measuring Soundscapes on resource-constrained devices

Nils KARGES[1]; Jeroen STAAB[2,3]; Jürgen RAUH[1]; Martin WEGMANN[1]; Hannes TAUBENBÖCK[1,2]

[1] Institute of Geography and Geology, University of Würzburg, Würzburg, Germany

[2] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Weßling, Germany

[3] Geography Department, Humboldt-University Berlin, Berlin, Germany

ABSTRACT

According to the WHO, noise is a growing problem in urban areas and the second most common environmental cause of health issues in Europe. As a complementary approach to noise maps based on sound pressure levels, soundscape maps can be a useful tool for urban planning, providing more information about how people perceive acoustic environments. This study describes an in-situ soundscape monitoring system based on WASN (Wireless Acoustic Sensor Network) for statistical spatial-temporal prediction of soundscapes in urban areas. Soundscape data on specifically defined spatial scales were observed and evaluated using a microcontroller with a 32-bit nRF52840 Nordic Semiconductors CPU and 1MB of memory in a multifunctional urban area. The use of TinyML enabled machine learning algorithms provided state-of-the-art soundscape classification to a low-cost edge device with extreme resource constraints regarding memory, speed, and lack of GPU support. Sound source types are classified into anthrophony, traffic, biophony, and geophony sounds using ESC-50 for evaluation. Our final MFCC-based CNN achieved an accuracy of 81.6% and even reached higher accuracy in the enclosed studio test. The results show that it is computationally feasible to classify soundscapes on low-power microcontrollers and potentially inform decision-makers based on extended sound analysis.

Keywords: Soundscape, TinyML, WASN

1. INTRODUCTION

The International Organization for Standardization (ISO) defines a soundscape as an "acoustic environment as perceived, experienced, and/or understood by a person or persons in a given context" (1). Like a fingerprint of the moment, variable in space and time, urban soundscapes contribute to the perceived quality of the urban environment and the identity of a city. Each urban area has a unique soundscape because of its multi-layered social and physical structure. There is a constant interplay between society and space, in which space configurations are profoundly intrinsic to human activity (2). The presence of ambient sounds can evoke thoughts and emotions, influence our mood, and thus guide our behavior (3–5). A listener trying to define a soundscape will always be influenced by the complex interactions of several variables, including physical, psychological, and physiological factors. Even minor variations in a few physical characteristics can cause similar acoustic environments to be perceived differently (6). This complexity of these phenomena makes modeling soundscapes a challenging task (7). Because of the potential positive impact that an appropriate acoustic environment can have on the well-being of citizens and the attractiveness of the city, the design of the acoustic environment of public urban spaces has received considerable attention for two decades (8–10). It is thus one indicator to address urbanization problems to meet the United Nations' demands for "sustainable cities and communities" and "good health and well-being"(11).

[1] nils.karges@uni-wuerzburg.de
[2] jeroen.staab@dlr.de
[3] jürgen.rauh@uni-wuerzburg.de
[4] martin.wegmann@uni-wuerzburg.de
[5] Hannes.Taubenböck@dlr.de

So far, noise mapping has mostly been based on sound pressure level (SPL) analysis. An example is the large-scale noise mapping application using publicly available data, context-aware feature engineering, and the linear land-use regression (LUR) model by Staab et al. (15). Mapping noise accurately and in its temporal variability remains a challenge. Beyond the standard noise maps demanded for large cities and highways in Europe, there is a need to measure soundscapes in greater detail. Therefore, we see our approach as a potential complement to maps based on sound pressure levels. Measuring soundscape empirically at the edge can be a valuable tool for urban planning, providing more information about how people perceive acoustic environments. The soundscape approach adds additionally a significant number of acoustic parameters to characterize the dynamics of the acoustic environment (see Appendix A in ISO 12913-2).

Traditional soundscape surveys are carried out by experts in-situ: the acoustic environments are identified, and the time and location of their appearance logged. Intensive research on soundscape indices in biodiversity and the emergence of soundscape ecology have given birth to new ideas on soundscape taxonomy (16,17). Krause (27) used the terms "biophony" and "geophony": the first term describes the complex arrangements of biological sounds and other ambient sounds and the latter describes the non - biological ambient sounds of wind, rain, lightning, etc. Later, Pijanovski et al. (16) extended the taxonomy with the "anthrophony" category, describing the sound created by human activities such as musical performance or oral conversation. In this work, the taxonomy of "anthrophony" is additionally extended with the sub-category of "traffic sounds" to provide opportunities to test the influence of traffic in urban environments.

In previous studies, the dynamics of the acoustic environment became evident, which led to a new impetus to continuously monitor the dynamic acoustic environments in cities (16,17). With the rapid development of the Internet of Things (IoT) (18–20), sensors and embedded processors are becoming smaller, cheaper, and with increasingly powerful computing capabilities. Solutions, which were previously difficult to achieve using small form factor and vast deployment, have now been made possible through a combination of IoT and cloud technologies (21,22). There have already been several attempts to use such smart devices to monitor and detect sounds on edge to reduce the bandwidth of data transmission to the central server and to maintain privacy (23,24). In order to characterize areas of a city and their sound environment, low-cost wireless acoustic sensor networks (WASN) are a cost-effective solution for centralized monitoring of the environment. Although various WASN techniques have been developed to increase bandwidth efficiency, sensor nodes are typically used to collect data, neglecting their computational capacity and the potential benefits of acting as a data center to process the data. The Term edge processing is used, when data computation is done directly on the smart sensor node or at the gateway to save battery power and ensure data privacy. It allows organizations to analyze critical information at the node level and shorten the time to detect anomalies (25,26).

In this paper, we propose a low-cost solution for a soundscape classifier with onboard audio classification, embedded in a wireless sensor network based on a secure protocol. We focus on the local computing capability of the sensor nodes on classifying soundscape variables. Furthermore, we test to what extent the sensors and their measured values are comparable in a studio environment. Based on our results, we discuss the influencing factors that need to be answered in future research to classify the soundscape variables more accurately.

2. SOUNDSCAPE SENSOR

In this paper we present a system architecture to classify urban environmental sounds consists of three layers: edge layer (sensing unit), fog layer, and cloud layer. As illustrated in Figure 1, the first part of the edge layer contains the sensor product described below, additionally equipped with a long-range (LoRa) transceiver to collect, analyze and ultimately send the data via LoRa communication to the IoT gateway. The Arduino Nano 33 BLE Sense, now acting as an edge device, brings the measured data via machine learning into the inference, which describes the process of running data points into a machine learning model to calculate an output. The second part describing the fog layer is operating with an ESP32 and a LoRa transceiver, representing the LoRa gateway connected to the Internet. The LoRa gateway is responsible for receiving the LoRa packets. The third layer of the IoT architecture is the cloud layer for cloud services and global storage as well as web application servers.

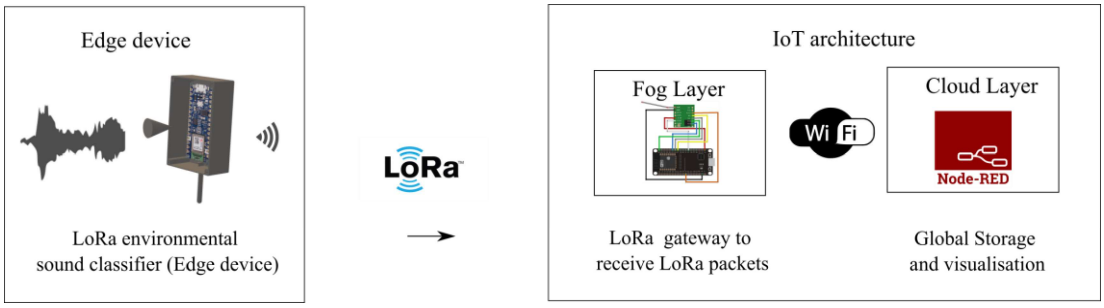

Figure 1 – The architecture of the soundscape monitoring system based on WASN.

2.1 Edge Device

There are infinite possibilities for different edge platforms (28), sensing units, and the software with which the platforms can be programmed. In this paper, the powerful core, in combination with very efficient energy consumption, high flexibility in terms of enabling/disabling modules, and the low material cost, justified the decision to use the Arduino Nano 33 BLE Sense as the systems development board and Figure 2 illustrating the individual modules attached to the development board. It includes several environmental sensors and the ability to run machine learning models with TinyML and TensorFlow™ Lite. Apart from the technical facts of different development boards, the support and availability of tested libraries are also a relevant consideration in hardware selection. We prefer to use development boards supporting a built-in digital signal processor module due to its significant increases in performance. The development board is further equipped with a MEMS microphone (MP34DT05), with omnidirectional sensitivity (–26 dBFS ±3 dB sensitivity) and a 64 dB signal-to-noise ratio which has a low power consumption and is capable of capturing and analyzing sound in real-time. Furthermore, the Arduino Nano 33 BLE Sense is equipped with the nRF52840 Nordic Semiconductors processor, a 32-bit ARM® Cortex™-M4 CPU running at 64 MHz. It has 1MB of program memory, and 256KB SRAM. The module size of the development board is 18x45mm (29).

The whole sensor module for environmental classification contains, besides the microcontroller, a RTC DS3231 I2C real-time clock which is required to get the correct time of the recordings. A 3V cell coin battery is added to retain the date and time while the power is off (e.g. when the main battery is empty). In addition, the RTC can be used to wake up the board from sleep mode.

Figure 2 – The deployable sensor module for environmental classification. Scale for reference.

Furthermore, Figure 2 shows the 11 cm high, 5 cm wide, and 4 cm deep waterproof plastic housing due to the TPE seal and a protection class of IP66/67 to secure the electronic components from moisture. The additionally attached conical periphery is filled with a wadding wind protector to dampen wind and noise. The development board also has a communications chipset equipped with a NINA-B306 chip, installed for Bluetooth communication. The chip uses version 5.0 and supports BLE. The NINA-B306 has an internal antenna and supports a maximum of 20 Bluetooth connections.

The BLE version of the soundscape sensor does not require external components except the battery and the SD-card module. The LoRa version of the proposed environmental classifier consists of the Nano 33 BLE Sense board, the NODEMCU ESP32 board, the LoRa RFM95 transceiver module in the 868 MHz band, a 6000 MAH lithium battery, and a 5 V voltage regulator. In this work, the NODEMCU ESP32 acts as an interface between the Nano 33 BLE Sense and the LoRa RFM95 transceiver module. ESP32 and Nano 33 BLE Sense are connected using UART serial communication protocol.

2.2 IoT architecture

The entire IoT architecture can be divided into three parts: The first part is the initialization of the system communication protocol, which is started by the soundscape sensor node. The second part contains the data handler (IoT gateway). This part is responsible for receiving the data packets sent from the edge device (soundscape sensor) and converting them from byte to float format. In this format, they can be sent to Node-Red in the last part of the IoT gateway application for analysis and visualization.

The developed system consists of a network of sensors forming a typical WASN, capable of retrieving data from the sensors and sending it to a cloud server. Another feature is the ability to communicate via LoRa peer-to-peer network using the RFM95 LoRa transceiver (30). This enables long-range spectrum communication, providing high-frequency immunity while maintaining low power consumption (31). The sensor node (soundscape sensor) describes the lowest level of our WASN, whose purpose is to collect data from the connected sensors and send the classified results to the IoT gateway. To accomplish these tasks, the sensor node requires a microcontroller and a set of sensors that can be permuted from node to node as needed for a specific solution. The sensor nodes consist of an ESP32, an RFM95 module, and the soundscape sensor for retrieving information which is then sent to the aggregation node.

The IoT gateway describes the solution to enable IoT communication. Here, the receiver unit consisting of ESP32 and RFM95 receives the data packets from the sensing node of the soundscape sensor and the RMF95. The receiver unit then sends the data to the cloud for global storage or visualization. The RFM95 LoRA transceiver module communicates with the ESP32 using the serial peripheral interface (SPI) communication protocol (32). Furthermore, the sensor data is sent via the IoT gateway to the IoT Cloud, where it is processed in real-time and visualized with Grafana. Grafana is a web-based visualization dashboard where the information must first be converted into the Node JS format using NodeRed. From NodeRed, the data is transferred via the API to the InfluxDB database, from where it can be queried into Grafana for further visualization. Although the technical setup is very complex, we endorse easy accessibility to the results for end users eventually.

3. METHODS

The proposed architecture attempts to solve the research problem of obtaining privacy in sound measurements in urban areas by moving the machine-learning processing from the cloud to the edge (24,33). The edge computing approach is being introduced to address the disadvantages based on cloud technology. As illustrated in Figure 3, the inference is executed directly on the microcontroller at the edge, allowing the system to operate in areas with poor or unstable Internet while protecting individuals' privacy. The sounds recorded by the microphone, which can include conversations between individuals, are processed on the microcontroller. The classified values alone are forwarded to the cloud layer to be processed by the end-user. Additionally, it has the advantage of reducing bandwidth, which enables the use of long-range but limited-bandwidth communication technologies such as LoRa (31).

3.1 Edge processing

The environmental classifier starts the inference by recording sound snippets with the microphone, which describes the first step in Figure 3. This recorded sound is passed to the feature extractor, which processes the signal as a Mel Frequency Cepstral Coefficient (MFCC). This part of the application is responsible for managing the sounds captured by the microphone and describes the input tensor of the model. The application of the acquisition unit describes the TensorFlow lite interpreter, which is directly connected to a TensorFlow Lite model. This part of the application is responsible for inference based on the environmental sound recordings captured by the microphone. Subsequently, the Convolutional Neural Network (CNN) model is integrated into the system as a C byte array, describing

the activity predictor, which determines the model's output and decides on a prediction based on thresholds. In the last step, the environmental classifier sends the data from the sensing unit to the system gateway via LoRa. After the device is set up, the loop function is called endlessly and is set to sleep mode when the classification results are successfully transmitted. If the transmission fails, the message will be resent up to five times and then stored on the SD card. The message is limited to 85 Bytes and consists of the header with an identifier, a timestamp, and the classification results (anthrophony, traffic, biophony, geophony).

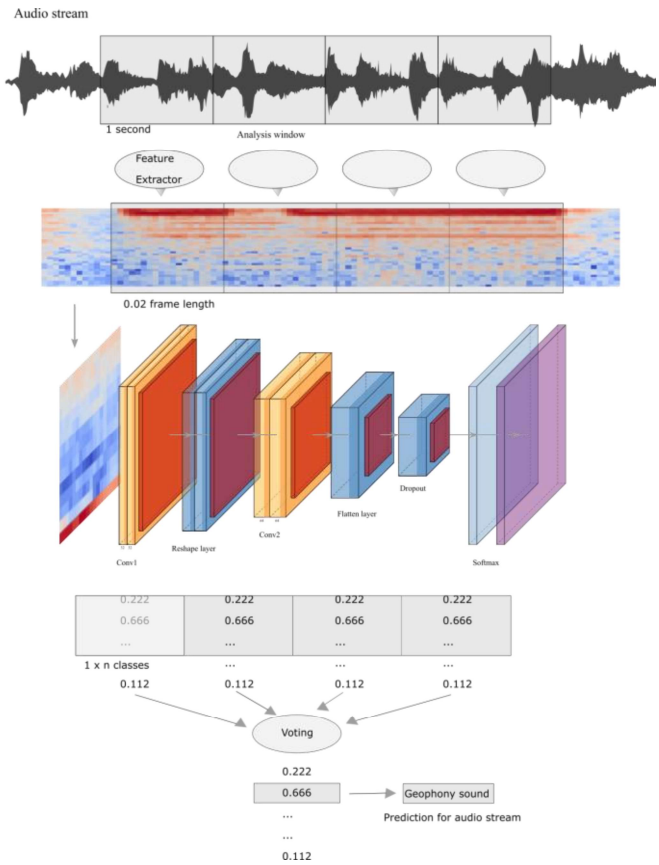

Figure 3 – Audio transmission process from detecting the signal to the voting for one of the four classification classes.

3.2 Dataset

To describe the machine learning process on edge, it is crucial to explain the dataset from which the model learns. Here, many datasets exist to identify environmental sounds that sample large numbers of audio with labels for each sample identifying which sound it is (34,35). Due to its versatile training data, the dataset used in this project is the ESC-50 from Piczak (36). The ESC-50 dataset includes a comprehensive summary of results on its Github repository with over 40 entries. The best models to date (July 2022) achieved 97.00% accuracy (37). The ESC-50 dataset from Piczak, suitable for comparing environmental sound classification methods, is a labeled collection of 2000 environmental audio recordings. The recordings in the dataset consist of 5-second recordings divided into 50 semantic classes, loosely arranged into five main categories. To organize the five major categories of the ESC-50 dataset into the taxonomies adapted from Krause and Pijanovski, the class "animals" is renamed biophony, and the class "natural soundscapes & water sounds" is renamed geophony. The categories "Human, non-speech sounds" and "Interior/domestic sounds" have been reclassified to the anthrophony class, representing human-generated sounds. The class of exterior/urban noises was empirically divided into the class of traffic sounds and anthrophony. On the one hand, traffic sounds can be understood as a subcategory of anthrophony sounds. On the other hand, traffic noise plays a unique and significant role in the urban context, so these sounds generated by vehicles (and thus by techniques) should be differentiated from such sounds that originate directly from humans. That is why we have designated traffic sounds as a separate category. The dataset now

consists of anthrophony sounds (As), geophony sounds (Gs), traffic sounds (Ts) and biophony sounds (Bs) as illustrated in Table 1.

Table 1 – After Piczak (2015) 5-second-long recordings organized into 28 semantical classes arranged into 4 major categories.

biophony Sounds	geophony Sounds	anthrophony Sounds	traffic Sounds
Dog	Rain	Footstep	Car sound
Insects (flying)	Sea waves	Laughing	Helicopter
Crow	Wind	Door knock	Siren
Sheep	Pouring water	Breathing	Car horn
Chirping birds	Thunderstorm	People talking	Train
Rooster	Crackling fire	Clapping	Airplane
Pig	Water drops	Coughing	Engine

3.3 Preprocessing

In the preprocessing, the signal in the form of a MFCC gets passed with a frame length of 0.02 seconds to the feature extractor. As shown in Table 2, the frame length, frame stride, FFT length, the number of coefficients, the number of filters, and the low to high-frequency band edge of the mel-scale filterbanks, as well as a local mean normalization of the signal, are extracted from the MFCC. Furthermore, different data augmentations like adding white noise or masking time- and frequency bands are performed to reduce the risk of overfitting due to data inadequacy and improve the efficiency and accuracy of the model.

Table 2 – Description of the individual parameters of the model.

Parameters	Values
Samplerate (Hz)	19000
Number of coefficients	13
FFT length (samples)	256
Frame lenght	0.02
Frame stride	0.01
Filter number	32
EpocAs	100
Learning rate	0.005
Training samples	2000
Validation samples	400

3.5 Training and performance

The training was performed for each individual analysis window, with each window having the associated label added to the audio clip (anthrophony, biophony, geophony, traffic). 75% of the sounds of the dataset was used as a training set, and 25% of the dataset were kept as a validation set. During training, audio clips are randomly selected from the four categories in the training set. For each audio clip, a time window is selected at a random position. In this way, the time-shifted data expansion is effectively implemented. To evaluate the model for the entire audio clip, an additional passover the validation set combines the predictions from multiple time windows. The learning rate used in this model was set to 0.005 and trained for up to 100 epochs. The training cycles were limited to 100 cycles for this test, as there is a risk that the data set is too tuned for the specific test set and might perform poorly with new data (overfitting). A summary of the experiment settings can be found in Table 2.

3.6 Evaluation

To evaluate the predictions of the resource-constrained device, the classifier is tested on two levels: First, the baseline is tested, where the training data is tested against the validation data. Second, live audio data is recorded from the built-in microphone and then classified. Subsequently, the selected

models are evaluated on the test set in each fold, giving us the baseline results used to compare against the results that our microcontroller will return once we run our model there. Further to the standard cross-validation for ESC-50, the model performance is also evaluated by separating foreground and background noise. Therefore, a confusion matrix is created in RStudio using the caret (38) and gmodels (39) package. The first confusion matrix is generated describing the baseline, calculated using the training data from the ESC-50. The second confusion matrix is derived from the result from the live audio-recording by the microcontroller.

Next, the trained models were tested on the development board with live audio on the microphone. To check the output of two sensors for similarity, both sensors are placed next to each other in front of two active near-field monitors (Adam A5X). Traffic sounds, anthrophony sounds, biophony sounds, and finally geophony sounds are played on the speakers in 5-minute sequences over a period of 1 hour. All soundscape sequences represent 1-hour recordings of youtube.com scenes according to the criterion of presenting the purest possible sound from the respective class. For example, a 1-hour soundscape with only traffic noise was used for the traffic sounds class, a sound file with rain and storm sounds was used for geophony sounds, and an hour full of bird sounds was used for biophony sounds. In contrast, the anthrophony class was represented with a long walla (reflecting the varying textures of crowd noise in the different countries) of people on the market side. In one hour, all four classes were repeatedly classified three times each. Here, the sensor is subjected to a test that simulates the measurement in the natural environment and plays soundscape recordings from the four classification classes for the sensor to evaluate the recordings. The experiment is shown in Figure 6 in the form of a plot divided into a dot plot in the lower part and a stripe plot in the upper part. Thus, the dots indicate the classification share of each recorded second, and the stripes describe the majority vote of each classification share per second. The resulting dataset serves as the basis for calculating the confusion matrix from Figure 5b).

4. RESULTS

4.1 Training performance

The loss diagram and the accuracy diagram in Figure 4 show that at the beginning of the first epochs, the model indicates comparable performance in both the training and the validation data sets. Subsequently, it can be seen that the parallel plots start to diverge more and more with more epochs. The accuracy of the training and validation data shows a similar tendency. The *val_accuracy*, indicating the accuracy of the predictions of a randomly selected validation data set after each training period, does not decrease but remains constant at about 80%, whereas the overall accuracy continued to increase. The overall loss decreases after almost every epoch and approaches 0. Like the *val_accuracy*, the *val_loss,* describing the value of the cost function for the cross-validation data, appears to be stagnant.

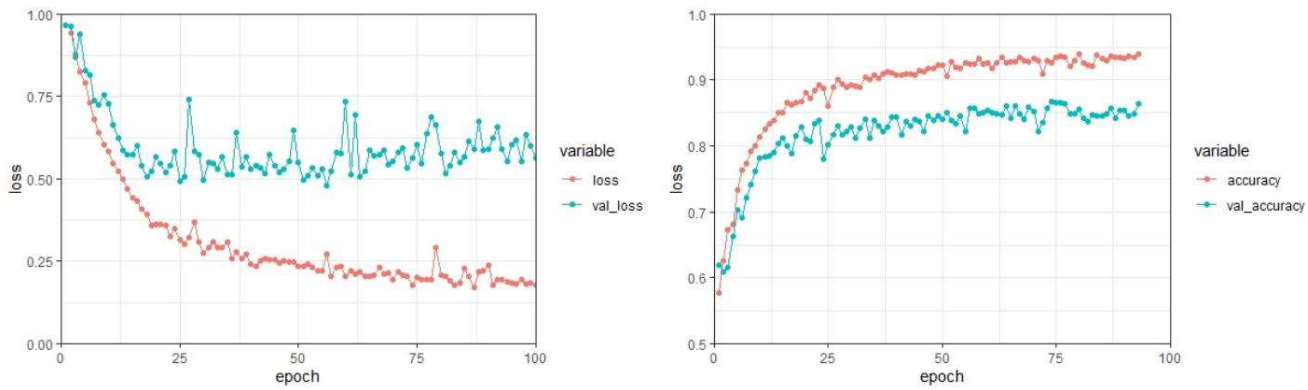

Figure 4 – Performance of the loss function of the training and validation datasets on the left and the accuracy of the train- and validation data on the right.

4.2 Evaluation

The confusion matrix of the baseline model shows the most accurately predicted classes were anthrophony sounds (As) (91.3%), traffic (Ts) (89.1%), followed by the average performance of the classes biophony sounds (Bs) (71.4%) and geophony sounds (Gs) (74.7%). In addition, 22.6% of the sounds labeled biophony and 13.6% of the geophony sounds were misclassified as anthrophony sounds. Furthermore, 9.9% of biophony sounds were misclassified as traffic sound and 9.4% of the traffic sounds were classified as anthrophony sounds. As described in Figure 5, we use the results of the baseline model (a)) to compare with the results of the sounds recorded in the test environment (b)).

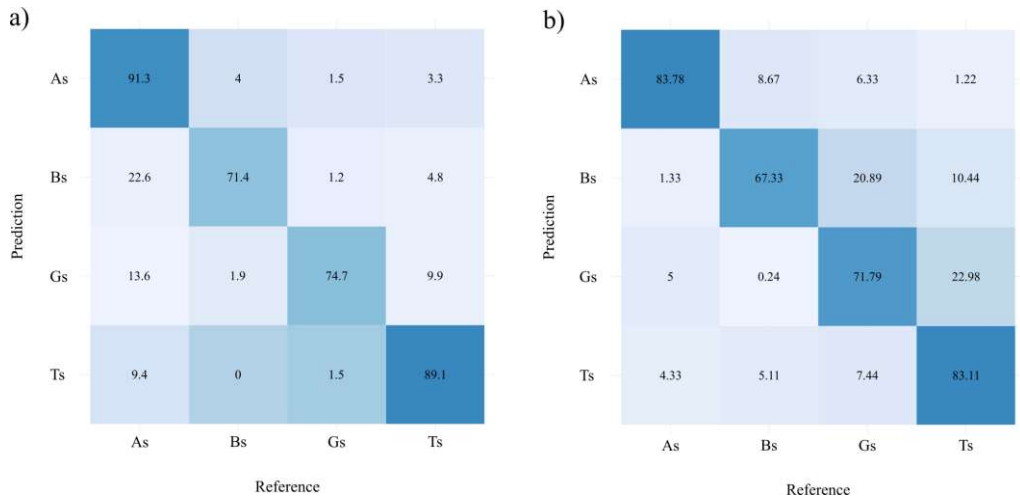

Figure 5 – a) The confusion matrix of the baseline model. b) The confusion matrix of the sounds recorded in the test environment. Displayed in percent (%).

As can be seen in Figure 5b), the class of the anthrophony sounds was best classified with 83.78%, closely followed by class traffic sounds with 83.11%. The class of the biophony sounds has a classification accuracy of 67.33%, representing a decrease in classification accuracy of 4.07% compared to the baseline confusion matrix. geophony sounds was classified with 71.79%. anthrophony sounds show a lower classification accuracy of 7.52% and a 5.99% decrease in the accuracy while classifying traffic sounds. The overall accuracy achieved by the baseline model is 81.6% and the model accuracy of the measurement in the test environment, achieved 76.50%. The similarity of the classification results is visible but still shows slight variations in the individual sequences. Some of the classification results show clear patterns in Figure 6, it can be seen that the classes are repeatedly confused with similar classes. In the stripes of the traffic sounds, it can be seen that, as shown in Figure 5b.), 22.98% of geophony sounds and 10.44% of biophony sounds are misclassified. The class of biophony sounds offers a relatively homogeneous classification accuracy and is only sometimes confused by misclassifications of anthrophony sounds with 8.67%. The class of geophony sounds is most affected by misclassifications. Here the most confusions occur in the area of biophony sounds with 20.89% and 7.44% for traffic sounds. Additionally, can be seen that most misclassifications appear in the lower classification share.

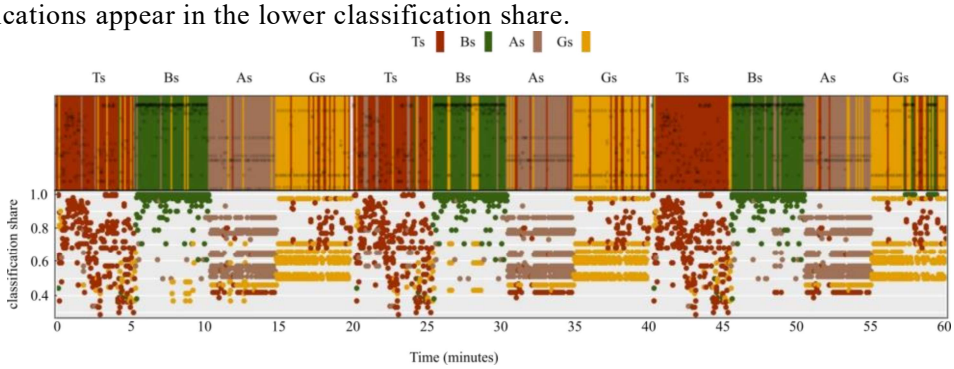

Figure 6 – The sensor's classification share per second.

4.3 Output

Once the ML-based soundscape classifier has been evaluated, the soundscape sensor can be deployed and described in terms of the data it provides. As shown in Table 3, the date, time, and geographical position (lon,lat) and the predictions from one of the four classes are placed in a separate column. The prediction value shown is referred to as the classification share. This data format is then received as a .txt file from the IoT gateway, and the metadata is transmitted via LoRaWAN to the cloud and visualized via Grafana. This empowers urban planners and gives insights into spatial-temporal soundscape characteristics of different urban areas.

Table 3 – Representation of the output format of the environmental classifier with the division of the predictions into the individual classes.

Date	Time	Lon	Lat	geophony share	biophony share	anthrophony share	traffic share
01.01.2022	05:00:00	9.565957	49.474098	0.09	0.14	0.54	0.22
01.01.2022	05:00:10	9.565957	49.474098	0.15	0.06	0.00	0.79
01.01.2022	05:00:20	9.565957	49.474098	0.12	0.04	0.28	0.56
…	…	…	…	…	…	…	…
01.01.2022	21:59:59	9.565957	49.474098	0.82	0.14	0.04	0.00

The Heatmap visualization in the monitoring dashboard displayed in Figure 7, shows an example of the change in soundscape classifications spread over a whole day (05:00- 21:59) at a random test site. The location describes a site of high entropy in terms of its land use and includes a park and a road intersection with a pedestrian crossing within a 250m radius. Here, the emerging classification share at the given time plotted on the x-axis is shown for each class. The heatmap graphic chosen in this paper is just one of many visualization options of this web service, and only there to show that the whole system can monitor Spatio-temporal features of the soundscape at multiple locations.

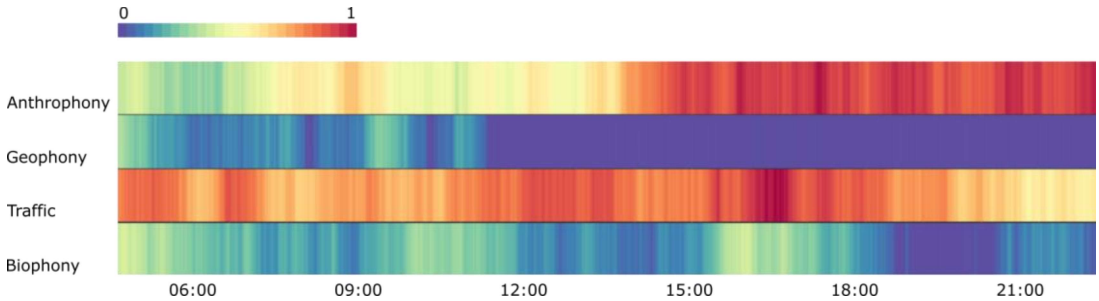

Figure 7 – Visualization of the data flowing into the cloud layer.

5. DISCUSSION

The methodology used and the goal we set, namely to classify the basic sources of a soundscape, work in principle. Still, there are some things to consider when evaluating the on-sensor classification process. On the one hand describing the model performance with the interaction between feature representation and RAM and CPU usage as well as the classification set-up itself. The reduced feature representation with a reduced number of predictions is used to minimize the peak RAM usage and flash usage so that the model can perform on the microcontroller. Compared to the SB-CNN by Salamon and Bello (40), the analysis window used is lower (1000 ms versus 1765 ms) and leads to a reduction of RAM and CPU usage. Furthermore, we keep the window increase low (250 ms), which decreases the performance, but also requires much less CPU usage.

In order to address potential measurement issues, the microphone itself has challenges: its placement and orientation, its spatial context as well as the temporal context of the measurements. Regarding the microphone, the acquisition circuit forms the subsystem for acoustic data acquisition. Here, the selection criteria for microphones are usually frequency range, flat frequency response and high sensitivity. The MP34DT05 is a Class 2 microphone where Class 1 microphones have a wider frequency range with less error tolerance then those of class 2. Furthermore, it is expected that the signal to noise ratio on the onboard microphone is lower compared to an external microphone. However, it is important to note that a trade-off between performance and cost is taken into account,

as one of the main requirements is to keep the cost low in order to be able to create large scale sensor networks (41).

A possible reason for the relatively low accuracy is the class imbalance of the ESC-50 dataset. Here, the anthrophony sounds as well as the traffic sounds predominate, and this is reflected in the confusion matrix of the model itself (Figure 5). We consider it crucial to design a training data set to analyze soundscapes to counteract underrepresented classes. The overall accuracy of the base model reached 81.6%, and the model accuracy of the measurement in the test environment reached 76.50%. The system proposed in this paper and the high classification results can support the soundscape analysis. Furthermore, the impact of the lower accuracies on application studies needs to be evaluated in future studies. The loss function in Figure 4 clearly shows convergence but is not overfitting. The decaying of the learning rate, shown in the form of the plateau of the validation set, could be due to the relatively high value of 0.05. When comparing the baseline model accuracy with the accuracy of the practical test application, a decrease in the performance given by the studio conditions are measured. We assume that this could be because there are fewer different sounds in the natural application area than the large number of sounds represented in the dataset. Additionally, the most representative sound examples used in the test dataset give an even more precise representation of the environmental sounds. It also needs to be mentioned, that the recordings of the respective classes sometimes also contain a class overlap. This is due to the fact that, on roads with predominantly traffic sounds, geophony sounds are usually also represented. Furthermore, as seen in the model, some classes like geophony and traffic sounds are often confused with each other because they serve similar frequencies, as seen in the spectral similarity. Based on these observations, there is likely a subtle similarity in their Log-Mel spectrogram features that the model cannot discriminate. The mixture of the individual classes in their respective weighting and background and foreground noises also appeared helpful in a soundscape characterization and are suggested to be analyzed in future studies.

6. CONCLUSIONS

Efficient methods for assessing and monitoring environmental sound diversity are central to urban research and critical to sound management. Technological advances in acoustic remote sensing are inspiring new approaches using soundscapes to understand and identify major ecological or human-oriented sound patterns by understanding the interrelationships of different sound sources such as animals, people, or the environment itself. Intensive descriptive methods capturing soundscapes are required for urban planning or design, with the ultimate goal of creating spaces of high acoustic quality.

Our proposed end-to-end IoT system combines machine listening at the edge with cloud services that provide real-time analytics of urban residential areas. At the edge, each sensor node performs sound classification. Due to the small size, low price, and low power consumption of LoRaWAN transmission in WASN, the presented soundscape sensors make it possible to extend its coverage and increase the Spatio-temporal density with limited rersources.

Further development potential is, e.g., to vary the granularity of the classification types at different locations to capture the likely predominant sound events at a given area more accurate. Various enhancements are currently being made to this system to improve its capabilities further. One of these enhancements is including a multi-label sound event localization and detection (SELD) system (42,43), as reported in the DCASE 2020 competition. Our proposed approach also provides customizable data visualization to monitor each sensor node in real-time. It includes retrospective descriptive analysis using aggregated metadata of the sound received by our sensor network. Although the quality of the data collected remains to be investigated, networks of these monitors placed at appropriate locations allow addressing a spatially dense soundscape monitoring across a city, ensuring a more reliable urban soundscape management with lighter installation and maintenance costs.

ACKNOWLEDGEMENTS

This study was partly funded by the German Federal Environmental Foundation (DBU). Furthermore, this research would not have been possible without the countless tools from the EAGLE Master program (University of Wuerzburg) used to construct the sensor.

REFERENCES

1. 14:00-17:00. ISO 12913-1:2014 [Internet]. ISO. [cited 2022 Jun 9]. Available from: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/21/52161.html
2. Lefebvre H, Nicholson-Smith D. The production of space. Vol. 142. Oxford Blackwell; 1991.
3. Axelsson Ö, Nilsson ME, Berglund B. A principal components model of soundscape perception. The Journal of the Acoustical Society of America. 2010 Nov;128(5):2836–46.
4. Hong JY, Jeon JY. Influence of urban contexts on soundscape perceptions: A structural equation modeling approach. Landscape and Urban Planning. 2015 Sep 1;141:78–87.
5. Non-auditory factors affecting urban soundscape evaluation: The Journal of the Acoustical Society of America: Vol 130, No 6 [Internet]. [cited 2022 Jun 10]. Available from: https://asa.scitation.org/doi/abs/10.1121/1.3652902
6. Kang J, Aletta F, Gjestland TT, Brown LA, Botteldooren D, Schulte-Fortkamp B, et al. Ten questions on the soundscapes of the built environment. Building and Environment. 2016 Nov 1;108:284–94.
7. Aletta F, Kang J, Axelsson Ö. Soundscape descriptors and a conceptual framework for developing predictive soundscape models. Landscape and Urban Planning. 2016 May;149:65–74.
8. Zhang M, Kang J. Towards the Evaluation, Description, and Creation of Soundscapes in Urban Open Spaces. Environ Plann B Plann Des. 2007 Feb 1;34(1):68–86.
9. Yu L, Kang J. Factors influencing the sound preference in urban open spaces. Applied Acoustics. 2010;71(7):622–33.
10. Herranz-Pascual K, Aspuru I, Iraurgi I, Santander Á, Eguiguren JL, García I. Going beyond quietness: Determining the emotionally restorative effect of acoustic environments in urban open public spaces. International journal of environmental research and public health. 2019;16(7):1284.
11. Nations U. Sustainable Development Goals [Internet]. United Nations. United Nations; [cited 2022 Jun 10]. Available from: https://www.un.org/en/sustainable-development-goals
12. Peckens C, Porter C, Rink T. Wireless Sensor Networks for Long-Term Monitoring of Urban Noise. Sensors. 2018 Sep;18(9):3161.
13. Zannin PHT, Ferreira AMC, Szerametta B. Evaluation of Noise Pollution in Urban Parks. Environ Monit Assess. 2006 Jul 1;118(1):423–33.
14. Chouard CH. [Urban noise pollution]. C R Acad Sci III. 2001 Jul 1;324(7):657–61.
15. Staab J, Schady A, Weigand M, Lakes T, Taubenböck H. Predicting traffic noise using land-use regression—a scalable approach. Journal of exposure science & environmental epidemiology. 2022;32(2):232–43.
16. Pijanowski BC, Villanueva-Rivera LJ, Dumyahn SL, Farina A, Krause BL, Napoletano BM, et al. Soundscape Ecology: The Science of Sound in the Landscape. BioScience. 2011 Mar 1;61(3):203–16.
17. Pijanowski BC, Farina A, Gage SH, Dumyahn SL, Krause BL. What is soundscape ecology? An introduction and overview of an emerging new science. Landscape Ecol. 2011 Nov 1;26(9):1213–32.
18. Mois G, Folea S, Sanislav T. Analysis of three IoT-based wireless sensors for environmental monitoring. IEEE Transactions on Instrumentation and Measurement. 2017;66(8):2056–64.
19. Lee HC, Ke KH. Monitoring of large-area IoT sensors using a LoRa wireless mesh network system: Design and evaluation. IEEE Transactions on Instrumentation and Measurement. 2018;67(9):2177–87.
20. Gómez JE, Marcillo FR, Triana FL, Gallo VT, Oviedo BW, Hernández VL. IoT for environmental variables in urban areas. Procedia computer science. 2017;109:67–74.
21. SONYC: a system for monitoring, analyzing, and mitigating urban noise pollution: Communications of the ACM: Vol 62, No 2 [Internet]. [cited 2022 Jun 13]. Available from: https://dl.acm.org/doi/abs/10.1145/3224204
22. Wong T, Watcharasupat KN, Lam B, Ooi K, Ong ZT, Karnapi FA, et al. Deployment of an IoT System for Adaptive In-Situ Soundscape Augmentation [Internet]. arXiv; 2022 [cited 2022 Jun 13]. Available from: http://arxiv.org/abs/2204.13890
23. Hassan AM, Awad AI. Urban transition in the era of the internet of things: Social implications and privacy challenges. IEEE Access. 2018;6:36428–40.
24. Atlam HF, Wills GB. IoT security, privacy, safety and ethics. In: Digital twin technologies and smart cities. Springer; 2020. p. 123–49.
25. Segura-Garcia J, Felici-Castell S, Perez-Solano JJ, Cobos M, Navarro JM. Low-Cost Alternatives for Urban Noise Nuisance Monitoring Using Wireless Sensor Networks. IEEE Sensors Journal. 2015 Feb;15(2):836–44.
26. Picaut J, Can A, Fortin N, Ardouin J, Lagrange M. Low-Cost Sensors for Urban Noise Monitoring Networks—A Literature Review. Sensors. 2020 Apr 16;20(8):2256.

27. Krause B. Anatomy of the Soundscape: Evolving Perspectives. JAES. 2008 Jan 15;56(1/2):73–80.
28. Chen J, Ran X. Deep learning with edge computing: A review. Proceedings of the IEEE. 2019;107(8):1655–74.
29. Nano 33 BLE Sense | Arduino Documentation [Internet]. [cited 2022 Jun 13]. Available from: https://docs.arduino.cc/hardware/nano-33-ble-sense
30. Firdaus R, Murti MA, Alinursafa I. Air quality monitoring system based internet of things (IoT) using LPWAN LoRa. In: 2019 IEEE international conference on internet of things and intelligence system (IoTaIS). IEEE; 2019. p. 195–200.
31. Bor M, Vidler JE, Roedig U. LoRa for the Internet of Things. In AUT: Junction Publishing; 2016 [cited 2022 Jun 28]. p. 361–6. Available from: https://eprints.lancs.ac.uk/id/eprint/77615/
32. Leens F. An introduction to I 2 C and SPI protocols. IEEE Instrumentation & Measurement Magazine. 2009;12(1):8–13.
33. Lv Z. Security of Internet of Things edge devices. Software: Practice and Experience. 2021;51(12):2446–56.
34. Salamon J, Jacoby C, Bello JP. A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia. 2014. p. 1041–4.
35. Cartwright M, Mendez AEM, Cramer J, Lostanlen V, Dove G, Wu HH, et al. SONYC urban sound tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network. 2019;
36. Piczak KJ. ESC: Dataset for Environmental Sound Classification. In: Proceedings of the 23rd ACM international conference on Multimedia [Internet]. New York, NY, USA: Association for Computing Machinery; 2015 [cited 2022 Jun 13]. p. 1015–8. (MM '15). Available from: https://doi.org/10.1145/2733373.2806390
37. Piczak KJ. karolpiczak/ESC-50 [Internet]. 2022 [cited 2022 Jul 4]. Available from: https://github.com/karolpiczak/ESC-50
38. Kuhn M. caret: Classification and Regression Training. Astrophysics Source Code Library. 2015 May 1;ascl:1505.003.
39. Warnes GR, Bolker B, Lumley T, Warnes MGR, Imports M. Package 'gmodels'. Vienna: R Foundation for Statistical Computing. 2018;
40. Salamon J, Bello JP. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Processing Letters. 2017 Mar;24(3):279–83.
41. Liu B, Towsley D. A study of the coverage of large-scale sensor networks. In: 2004 IEEE International Conference on Mobile Ad-hoc and Sensor Systems (IEEE Cat No04EX975). 2004. p. 475–83.
42. Nguyen TNT, Jones DL, Gan WS. Ensemble of Sequence Matching Networks for Dynamic Sound Event Localization, Detection, and Tracking. In: DCASE. 2020. p. 120–4.
43. Tan EL, Karnapi FA, Ng LJ, Ooi K, Gan WS. Extracting Urban Sound Information for Residential Areas in Smart Cities Using an End-to-End IoT System. IEEE Internet Things J. 2021 Sep 15;8(18):14308–21.