Environmental Science and Ecotechnology 22 (2024) 100450

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Estimating rainfall intensity based on surveillance audio and deep-learning

Meizhen Wang [a, b, c], Mingzheng Chen [a, b, c], Ziran Wang [a, b, c, d], Yuxuan Guo [a, b, c], Yong Wu [e], Wei Zhao [a, b, c], Xuejun Liu [a, b, c, *]

[a] Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing, 210023, China
[b] State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing, 210023, China
[c] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, 210023, China
[d] School of Information Engineering, Nanjing Normal University Taizhou College, Taizhou 225300, China
[e] Institute of Geography, Fujian Normal University, Fuzhou, 350000, China

ARTICLE INFO

Article history:
Received 21 September 2023
Received in revised form
4 July 2024
Accepted 5 July 2024

Keywords:
Surveillance audio
Rainfall intensity
Dataset
Regression
Deep learning

ABSTRACT

Rainfall data with high spatial and temporal resolutions are essential for urban hydrological modeling. Ubiquitous surveillance cameras can continuously record rainfall events through video and audio, so they have been recognized as potential rain gauges to supplement professional rainfall observation networks. Since video-based rainfall estimation methods can be affected by variable backgrounds and lighting conditions, audio-based approaches could be a supplement without suffering from these conditions. However, most audio-based approaches focus on rainfall-level classification rather than rainfall intensity estimation. Here, we introduce a dataset named Surveillance Audio Rainfall Intensity Dataset (SARID) and a deep learning model for estimating rainfall intensity. First, we created the dataset through audio of six real-world rainfall events. This dataset's audio recordings are segmented into 12,066 pieces and annotated with rainfall intensity and environmental information, such as underlying surfaces, temperature, humidity, and wind. Then, we developed a deep learning-based baseline using Mel-Frequency Cepstral Coefficients (MFCC) and Transformer architecture to estimate rainfall intensity from surveillance audio. Validated from ground truth data, our baseline achieves a root mean absolute error of 0.88 mm h^{-1} and a coefficient of correlation of 0.765. Our findings demonstrate the potential of surveillance audio-based models as practical and effective tools for rainfall observation systems, initiating a new chapter in rainfall intensity estimation. It offers a novel data source for high-resolution hydrological sensing and contributes to the broader landscape of urban sensing, emergency response, and resilience.

1. Introduction

Rainfall information plays a pivotal role in urban hydrology systems and finds wide-ranging applications across numerous research domains, including meteorological forecasting, water resource management, agricultural security, and urban planning [1–7]. The existing literature highlights the crucial significance of high-resolution rainfall data [8]. Due to climate change, rainfall in complex areas (such as urban areas and mountains) is becoming more varied and frequent. Therefore, high-quality rainfall data are always desirable. The relatively small scale of urban catchments and the specific demands of hydrological applications, particularly those requiring real-time insights, necessitate rainfall information with spatial resolutions as precise as 1 km and temporal resolutions of 1 min [2,9–11]. Nevertheless, current rainfall observation methods, such as ground- and satellite-based systems, still fall short of meeting these stringent resolution requirements [12]. Consequently, substantial efforts have been devoted to enhancing the resolution and precision of rainfall estimation. Leveraging the extensive utilization and technological advancements in surveillance systems, surveillance camera-based rainfall estimation (SCRE) has emerged as a promising avenue for achieving low-cost, high-resolution rainfall observations [7,13–15].

Most studies concerning SCRE often focus on surveillance video-

* Corresponding author. Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing, 210023, China.
E-mail address: liuxuejun@njnu.edu.cn (X. Liu).

https://doi.org/10.1016/j.ese.2024.100450

based rainfall estimation (SVRE) methods [4,14,16–23]. However, challenges remain in this area, especially under uncontrolled noisy conditions characterized by factors such as shading and lighting variations. As a result, there has been growing interest in surveillance audio-based rainfall estimation (SARE), aiming to alleviate the limitations of visual observation during environmental fluctuations. SARE has distinct advantages in the acoustic domain, including its all-weather capability, resilience to obstructions, and low data volume [7,13]. However, it is essential to note that SARE does not displace visual methods but complements rainfall estimation using acoustic signals. Therefore, effectively leveraging audio data from surveillance resources can advance SCRE development, offering novel perspectives and approaches to enhance current rainfall observation systems.

Early attempts at acoustic-based rainfall observation focused on rainfall events above the ocean or water area [24,25]. Various underwater acoustic sensors have been designed to record rainfall sounds generated by raindrops colliding with the water surface, and the acoustic characteristics, raindrop size distribution, and wind have subsequently been analyzed [26]. These studies claim this approach is an efficient way to observe rainfall. However, the differences in the acoustic environments between land and ocean pose challenges in transferring methodologies from oceanic to terrestrial settings. Hence, certain researchers have directed their efforts toward gathering acoustic data and formulating observation methodologies for rainfall, specifically in terrestrial environments. For example, Dunkerley proposed an approach to get unbiased rainfall duration and intensity data from tipping-bucket rain gauges using synchronized acoustic recordings to get rainfall start and end times [27]. Nevertheless, the majority of current studies necessitate specialized acoustic sensors [28,29], and/or can only approximate broad categories of rainfall levels by using a thresholds [30–34], decision tree [35], random forest [36], convolution neural network (CNN) [7,29,37,38], and parallel network [13]. Consequently, these studies have limitations regarding both high installation costs and the granularity of rainfall data. In 2017, Bedoya et al. proposed a linear regression approach for estimating rain intensity based on power spectral density features extracted from audio recordings in forested environments [31,33,34]. However, the general applicability of their method warrants further investigation, as audio features in forested areas may differ significantly from those recorded in urban environments, such as differences in underlying surfaces, background sounds, and weather conditions. In addition, although current advances in audio-based rainfall detection and classification are predominantly data-driven, and many authors have assembled their datasets, the availability of online datasets remains limited, with only the following datasets available online: Audio/Video Database for Rainfall Classification (AVDB-4RC)[1] [39], RAZ[2] [7], and audio-extreme rainfalls dataset[3] [29]. Unfortunately, these datasets suffer from the shortcomings of being small in size (less than 5000 sound slices) and having limited diversity (only rainfall levels are labeled). Moreover, the AVDB-4RC dataset names its audio files based on respective rainfall intensity values; however, it comprises only 15 rainfall audio samples, each lasting 22 s. These recordings were captured using a microphone shielded by a plastic shaker at a singular location, resulting in limited variations concerning rainfall events and underlying surfaces within this dataset.

Overall, challenges persist in SARE concerning the granularity of rainfall estimation, methodological adaptability, and dataset volume and diversity, leading to limitations in generating finely detailed rainfall data in current studies. Therefore, a publicly available dataset and practical approaches have yet to be developed to utilize surveillance audio networks for calculating rainfall intensity. This paper introduces the Surveillance Audio Rainfall Intensity Dataset (SARID) and a baseline model to overcome existing limitations and provide a better benchmark for SARE approaches. SARID provides a comprehensive and diverse data collection for surveillance audio-based rainfall intensity estimation (SARIE). Through deep learning, the baseline model establishes the relationship between rainfall intensity and surveillance audio.

We developed a data collection pipeline with thorough annotation to build an effective dataset for data-driven training and validation. First, we collected surveillance audio and corresponding meteorological data. Each audio recording was then segmented and annotated with the following: (1) rainfall intensity (RI), (2) meteorological information (i.e., temperature, humidity, air pressure, and wind speed), (3) underlying surface data, and (4) background noise details (e.g., car sounds, human activity, and animals). To understand the complexity of SARID, we conducted basic experiments. Using common acoustic features and deep learning frameworks, we created a regression model that maps surveillance audio to RI. These experiments highlight the effectiveness of the optimal baseline model by using the Mel-Frequency Cepstral Coefficient (MFCC) as the input and the Transformer [40] as the network in SARE. Eventually, this study contributes to developing a comprehensive dataset and provides an effective and low-cost SARIE system that was previously only conceptual. The study outcome is expected to enhance the feasibility of using surveillance camera networks for hydrology sensing and to provide valuable insights for future endeavors.

The main contributions of this paper are as follows.

(1) SARID, the only open audio-based rainfall intensity dataset, provides surveillance audio recordings of rainfall events in urban areas with extensive annotations. This dataset is highly advantageous for training, validation, and data-driven analysis.
(2) We develop and analyze an effective baseline model for SARIE using the MFCC input and the Transformer network. The experimental results demonstrate that the model is superior to other machine learning methods.

The rest of the paper is organized as follows: Section 2 details the dataset generation process. Section 3 introduces the baseline model. Section 4 validates the model and discusses the system's effectiveness. Section 5 provides the conclusions of the study.

2. Dataset

SARID is used to evaluate rainfall estimation methods based on surveillance audio. SARID must include a wide range of variability in rainfall intensity, background noise, weather conditions, and underlying surfaces to achieve this goal. In addition, it is critical that SARID adheres to a standardized annotation format and is scalable to allow for the addition of more data in the future. To ensure the suitability of the dataset for accurate training and evaluation, the audio annotations must maintain consistency, accuracy, and completeness for the specified rainfall events. This section outlines the procedures for collecting and annotating SARID to meet these objectives.

[1] https://github.com/vicosystems/AVDB-4RC.
[2] https://pan.baidu.com/s/1MTv8cbLLh1sB3yEatNg8mw.
[3] https://pan.baidu.com/s/1DcV8ei73SWa3klrP-5Jbxg (Code: b48n).

2.1. Data collection

2.1.1. Collection devices and distribution

For SARID, we obtained all audio recordings and rainfall information from surveillance cameras and a meteorological station on the Xianlin Campus of Nanjing Normal University, China. The meteorological station has various environmental monitoring instruments, including a disdrometer and temperature and humidity sensors. It provides nearly real-time updates of weather conditions, such as rainfall intensity, temperature, and humidity, every minute, meeting the specified temporal resolution requirement of 1 min, as mentioned earlier. To fulfill the spatial requirement of 1 km, all surveillance cameras were strategically positioned within 500 m of the weather station (Fig. 1). Given the complexity of the meteorological station parameters, we refrain from providing an exhaustive description here. Table 1 displays the key specifications of the surveillance camera used for data collection (Ezviz *C*S-CB3).

2.1.2. Data collection procedure

With the data collection devices selected, the next critical step was to assemble a carefully selected set of candidate audio recordings. To strengthen the data representation, SARID diligently collected surveillance audio samples from various underlying surfaces and categorized them into five different classes: road (concrete), road (marble), road (wood), urban meadow, and water. Within the road (concrete) category, we identified two distinct scenarios: one located within the campus grounds (Fig. 1a) and the other located at the campus entrance (Fig. 1b). Notably, the campus entrance experiences increased vehicular traffic, resulting in the amplified presence of background noise in the surveillance audio recordings. The road (wood) data were collected on an open-air terrace surrounded by nearby buildings (Fig. 1c). Conversely, urban meadow provided a softer surface (Fig. 1d). For the road (marble) category, we focused on surfaces composed primarily of regularly shaped marble tiles, which are known for their relatively uniform and flat composition (Fig. 1e). During data collection near water, often in remote locations, the recorded audio had minimal background noise (Fig. 1f). Due to the limitation of land use types around the meteorological station, the number of classes was limited to five. The extension of more classes will be considered in future iterations.

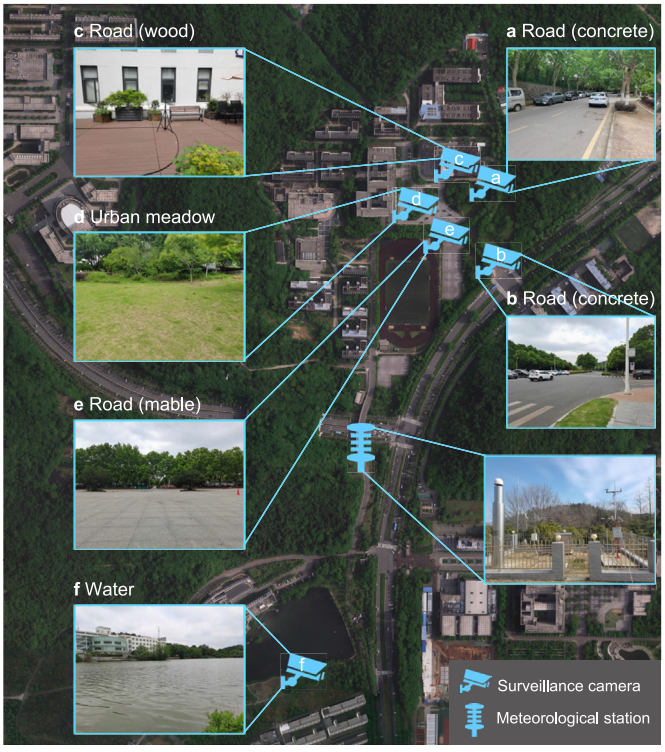

Fig. 1. Distribution of collection devices and different underlying surfaces.

SARID collected audio recordings from surveillance cameras capturing rainfall events on these substrates. Six audio recordings of rainfall events were secured, ranging from a minimum recorded RI of 0.04 mm h^{-1} to a maximum RI of 14.72 mm h^{-1}. The full details of these six rainfall events are provided in Table 2. Notably, the "time" series in Table 2 denotes the start and end times of the camera recordings on the respective days. Due to intermittent rain-free periods, only audio recordings of actual rainfall were retained. Following this initial audio recording phase, synchronized meteorological station data were collected to facilitate the preparation of the data annotations.

2.2. Data annotation

2.2.1. Audio annotated attributes

For evaluating estimation methods, the audio annotations include the following attributes for each rainfall event: (1) Rainfall information: This includes RI; (2) Scenario information: This includes environmental factors such as temperature, humidity, barometric pressure, and wind speed. It also includes details about the underlying surface and labels for background sounds (e.g., car, people).

An overview of all annotated attributes in SARID is provided (Fig. 2) In addition to the basic RI data, each sound recording in SARID is enriched with numerous scenario-related annotations. Variations in environmental factors can significantly affect raindrop characteristics, potentially influencing the generation and propagation of rain sound [41]. For example, changes in wind speed can alter the speed and direction of raindrops, thereby affecting the path and intensity of rain sound propagation. In addition, variations in temperature and air pressure can affect the size and evaporation rate of raindrops, further affecting rain sound production. As a result, atmospheric conditions were included as part of the annotation attributes. In addition, due to the inherent complexity of urban environments, surveillance audio often captures a variety of other sounds, such as vehicular traffic, human activity, and animal noises, which typically contribute to the acoustic landscape of rain sound. As a result, additional urban sounds classified as background noise were introduced. These annotations can be valuable for noise-filtering algorithms, ultimately improving the performance of audio-based rainfall estimation. To accomplish this task, each audio recording underwent a manual inspection with auditory analysis, and the identified noise was meticulously labeled in a CSV file. In the following section, we will return to this point when discussing the annotation procedure.

2.2.2. Audio annotated procedure

Given the need to annotate both existing and future audio recordings, designing a cost-effective yet high-quality annotation pipeline was imperative. The annotation pipeline can be divided into two primary tasks: data matching and data organization (Fig. 3). Note that surveillance audio is derived from surveillance video data, and due to the specific data storage format of the Ezviz surveillance cameras, an additional step of converting the original video data was required. Therefore, we manually reviewed all

Table 1
Key Parameters of surveillance camera.

Category	Parameter	Details
Model/Version	Model	CS-CB3
	Version	V100-2D2FL4GT
Camera	Sensor type	1/2.9 Progressive Scan CMOS
	Lens	2.8 mm@F2.0, visual diagonal angle 128°, horizontal angle 108°, vertical angle 56°
Interface	Audio input	Built-in high-sensitivity microphone
	Audio output	Built-in high-power speaker
Compression Standards	Video compression standard	H.265
	Video compression bitrate	Adaptive bitrate
	Audio compression bitrate	Adaptive bitrate
Image	Maximum image size	1920 × 1080
	Frame rate	Max: 15 fps
	Supported protocols	Ezviz cloud private protocol

Table 2
Description of recorded rainfall events.

ID	Rainfall event	Time	Primary underlying surface	Minimum RI (mm h^{-1})	Maximum RI (mm h^{-1})
1	2022-09-15	17:56–23:56	Road (marble)	0.04	3.9
2	2022-10-05	13:25–17:56	Road (wood)	0.04	14.72
3	2022-10-06	09:45–22:59	Road (concrete)	0.25	7.96
4	2022-10-08	23:50–23:59	Urban meadow	0.12	0.54
5	2022-10-26	15:37–23:48	Water	0.04	5.21
6	2022-11-21	22:37–23:59	Road (wood)	1.6	9.59

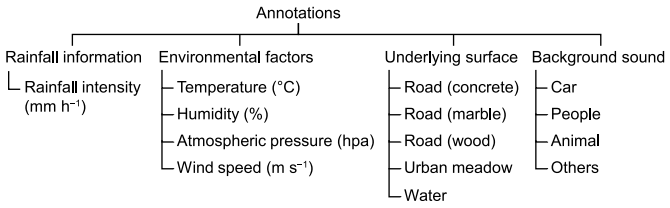

Fig. 2. SARID annotations.

original videos to obtain the start timestamp and duration, then converted them to a standard video format using "moviepy[4]", a Python library for video data processing. Subsequently, the corresponding audio was extracted.

The first task in annotating SARID was to match audio and meteorological station data based on their timestamps. For long audio files and meteorological station data, segments with non-zero RI were extracted by parsing the meteorological data. Then, using the temporal information, the corresponding audio segments were matched and divided into shorter audio files, typically 60 s in duration, although shorter durations were occasionally used for specific boundary data. After this initial matching stage, 1093 individual recordings were acquired. Next, the filename was used to label the annotation attributes of each audio recording. For example, consider the annotation information contained in the name of a particular audio file: "2022-09-15 17-56-00_0.19_22.335_92.35_2.491_0.892_ hiv00013_60_road (concrete). mp3". This filename information can be segmented using the "_" delimiter as follows.

- "2022-09-15 17-56-00": Time tag indicating that the recording time is September 15, 2022, at 17:56.
- "0.19": Rainfall intensity in millimeters per hour during the specified time interval.
- "22.335": The average temperature in Celsius for the given time period.
- "92.35": The average humidity observed during the recorded time interval, expressed as a percentage.
- "2.491": The average atmospheric pressure in hPa during the given time period.
- "0.892": The average wind speed in meters per second recorded during the given time period.
- "hiv00013": The original video file is associated with the current audio segment.
- "60": The total duration of the audio file, which is 60 s.
- "road (concrete)": Specifies that the underlying surface for the recording is a concrete road.

These detailed annotations provide comprehensive insights into each audio file, including the corresponding meteorological conditions recorded at specific times. After completing the above-mentioned task, a verification step was performed on the segmented instances to ensure the accuracy of the dataset. Since verifying annotations is a considerably time-consuming process and audio identification can be challenging, we used a random sampling approach. We selected some video files (about one-tenth of the original video files) and cross-referenced the timestamp information in the videos to verify the correctness of the annotations. In the next phase, we performed manual annotations of the noises identified within each audio segment. By listening carefully to each recording, we meticulously documented the exact start and end timestamps for each identified sound and assigned an appropriate category (e.g., "people") to each noise source. This process resulted in a separate noise annotation file in CSV format. Specifically, the annotation file contained the following fields: "n_start" (noise start timestamp), "n_end" (noise end timestamp), and "n_type" (noise type).

2.3. Dataset statistics

As part of the research focused on rainfall estimation using surveillance audio, we created a subset of shorter sound segments. Based on existing literature, previous studies have shown that

[4] https://github.com/Zulko/moviepy.

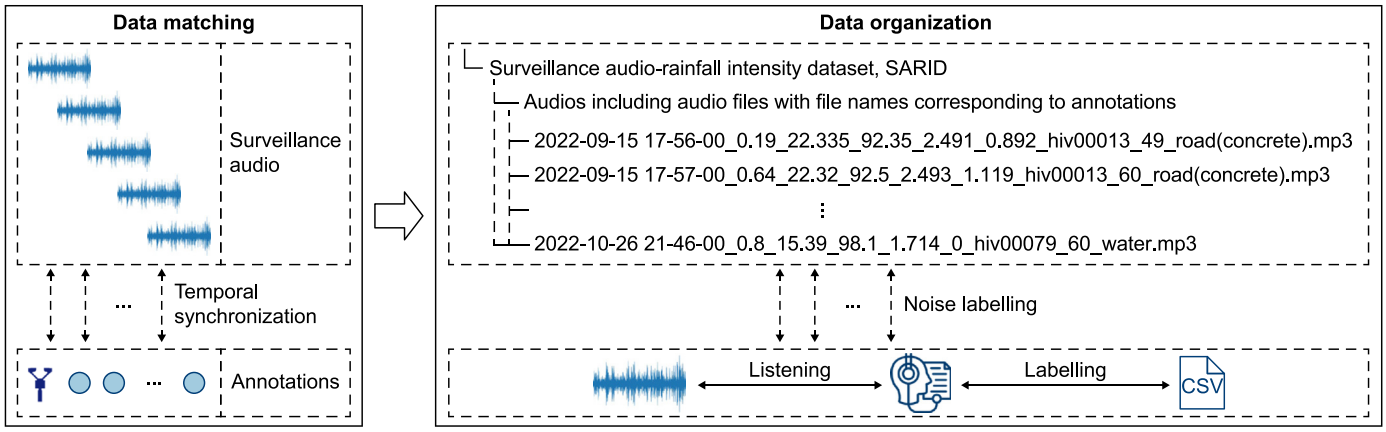

Fig. 3. Annotation pipeline.

providing subjects with 4-s snippets of environmental sounds can yield an impressive 82% accuracy rate for sound identification [42]. Therefore, we adopted this recommended duration as the maximum limit and proceeded to divide the audio files into 4-s slices. Furthermore, since the primary goal of this paper is to contribute a new dataset and develop a functional baseline model, the effect of background noise will be ignored in subsequent experiments. This resulted in a total of 12,066 labeled slices.

The SARID statistics, particularly for the subset used in subsequent experiments (as described in Section 3.3), are shown (Fig. 4). The distribution of audio samples per RI interval, with intervals of 0.5 mm h^{-1} (Fig. 4a). Note that these counts are plotted on a logarithmic scale. The interval labeled [0.5, 1.0] was the most common and contained 2562 sound slices. Rainfall event information and the number of slices corresponding to different underlying surfaces are shown in Fig. 4b and c. Among the recorded rainfall events, the underlying surface labeled "road (wood)" had the widest range of rainfall intensities, reaching a maximum of 14.68 mm h^{-1}. On the other hand, the "urban meadow" had the narrowest range, at only 0.42 mm h^{-1}. In terms of sample distribution, approximately 59% of the audio samples corresponded to "road (concrete)," while 18% represented "water." In addition, "road (marble)" accounted for 15% of the samples, followed by "road (wood)" with 7%, and "urban meadow" with 1%. The distribution of noise in the dataset is shown in Fig. 4d. It is important to note that the noise labels are relative to the original audio segments due to the fixed length of 4 s for our audio samples. A total of 2142 noise audio segments were labeled into seven classes, namely "car passing," "car whistle," "wind," "people," "animal," "other," and "hybrid." The "other" class represents cases in which the annotators could not identify a specific noise class, which is not uncommon due to the complexity of the surveillance soundscape. After filtering out segments with noise, 2527 rainfall audios ("no_noise") remained and were used to generate 4-s slices. Statistical information on the recorded meteorological data is depicted using a box plot (Fig. 4e). The data were normalized to facilitate visualization since the units of different meteorological elements vary. Temperature exhibited relatively large fluctuations, while humidity, barometric pressure, and wind speed exhibited more stable values. Because the recorded rainfall events mostly fall within the same season, with little variation in meteorological conditions, further analysis will require an expansion of the dataset to explore the role of meteorological information.

3. Baseline model

This section introduces the proposed baseline model, training methodology, and evaluation experiments. The baseline model has two primary objectives: to establish an effective framework for estimating RI based on surveillance audio and to select the most

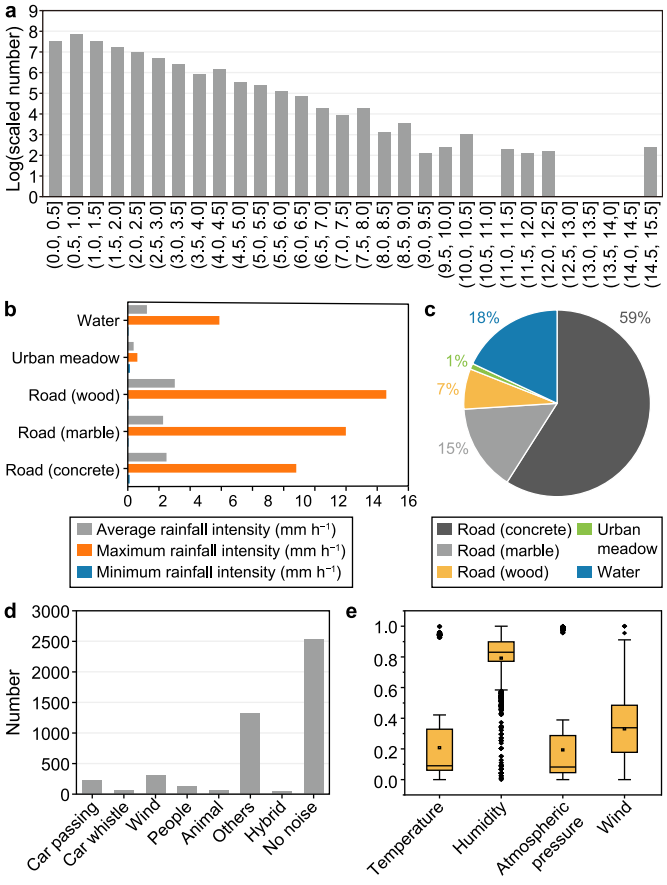

Fig. 4. Statistics of the SARID. **a**, Distribution of monitoring rainfall audio slices with a logarithmic scale. **b**, Information about rainfall events on different underlying surfaces. **c**, Number of audio slices per underlying surface. **d**, Number of annotated instances per sound category. **e**, Statistics of meteorological information.

suitable feature engineering techniques and network architectures for this challenge from existing, well-established research.

3.1. Baseline model overview

The basic model consists of three main components. The first component serves as the input to the model. Three commonly used acoustic features, namely Mel-spectrogram (Mel), Mel-Frequency Cepstral Coefficient (MFCC), and Short-Time Fourier Transform (STFT), were selected to represent the rainfall audio and used in subsequent experiments [7,13,43]. Mel captures the frequency content of audio signals and is widely used in audio signal processing. MFCC contributes crucially to capturing the spectral characteristics of audio signals, effectively representing various audio-related tasks, such as speech and environmental sound analysis. The STFT is employed to analyze the time-frequency domain of the audio signal, decomposing it into its frequency components over short, overlapping time intervals. These features were extracted using the Python library "librosa" [44]. The Mel coefficient and MFCC were set to 128, resulting in feature vectors with dimensions of 128 $\times$ 173. The STFT used default coefficients, resulting in dimensions of 1025 $\times$ 173 $\times$ 1.

The second component is a deep learning network that extracts feature maps from the input acoustic signal. Three widely used deep learning architectures suitable for audio data were analyzed: CNN, Long Short-Term Memory (LSTM), and Transformer. CNN is known for spatial modeling, LSTM is for temporal modeling, and Transformer uses self-attention to extract features at different levels. Note that in this configuration, only the "encoder" module of the Transformer serves as a feature extractor for the baseline network. Notably, both the input acoustic features and the deep learning network are single-function, meaning that a combination of a single input feature class and a single network class was used (e.g., using MFCC as the input feature and Transformer as the network structure), rather than using multiple features or networks. The network then feeds the feature map output from the last fully connected layer for rainfall intensity prediction. The entire baseline model follows a regression framework for estimating RI.

3.2. Baseline model structure

3.2.1. CNN-based baseline

In the CNN-based baseline model, similar to its application in image tasks, the convolution kernel spans the entire channel dimension while maintaining a limited range along the width dimension. This approach allows the CNN to capture patterns across the acoustic spectrum effectively while preserving the local temporal context (Fig. 5a). The network architecture consists of the following components: three convolutional layers with ReLU activation and batch normalization, three max-pooling layers with dropout, one global average pooling (GAP) layer, and two fully connected layers. After the CNN extracts the feature maps, the GAP layer transforms each feature into a fixed-size feature vector, fed into fully connected layers for RI estimation.

3.2.2. LSTM-based baseline

In the LSTM-based baseline model, the LSTM is used as the primary feature extraction structure, and the fully connected layers are used to construct a regression model for RI estimation. Unlike recurrent neural networks (RNNs), this model relies on the LSTM as a more effective temporal feature extractor. The network architecture (Fig. 5b) includes two stacked LSTM layers of 256 hidden units each. Fully connected and pooling layers exist to downscale temporal features and perform RI regression.

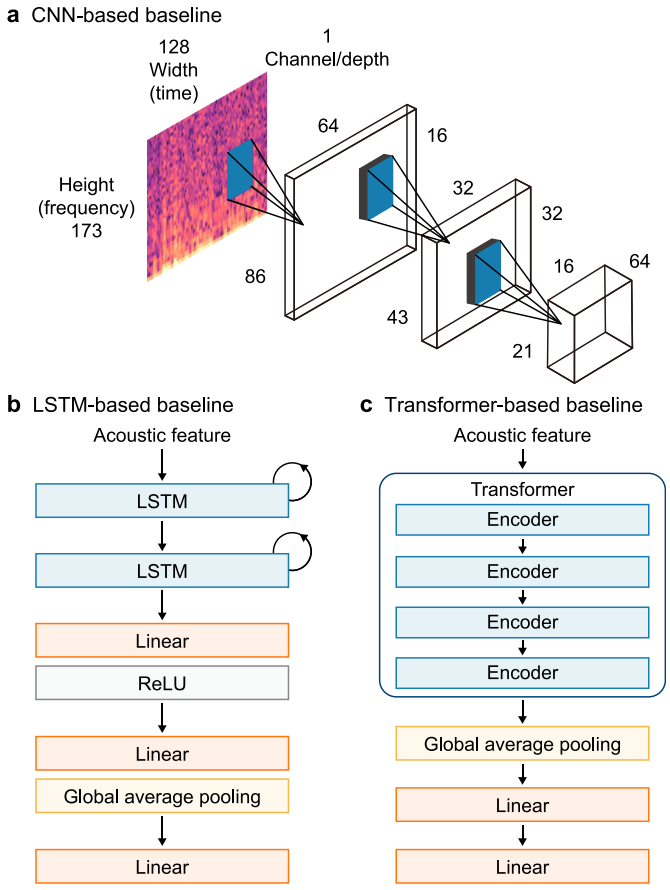

Fig. 5. Baseline structure. **a**, CNN-based baseline. **b**, LSTM-based baseline. **c**, Transformer-based baseline.

3.2.3. Transformer-based baseline

The Transformer-based baseline model employs the "encoder" portion of the Transformer as its fundamental structure to establish a feature extraction network (Fig. 5c). In this setup, the Transformer module comprises four stacked encoders with specific hyperparameters: the number of attention heads ("nhead") is set to 4, and the dimension of the feed-forward fully connected layer is configured as 512. The feature dimension extracted by the Transformer encoder remains consistent with the original. Then RI estimation is performed using a fully connected layer after obtaining a two-dimensional feature vector through GAP. This sophisticated utilization of the Transformer's "encoder" ensures the model can capture intricate patterns within the surveillance audio data for accurate RI estimation.

3.3. Training

The baseline model enables end-to-end training on SARID for single-forward RI estimation. The training process involves the selection of appropriate loss functions to optimize estimation performance, along with specific training strategies. The estimation loss, as defined in equation (1), uses a smooth L1 loss [45] that computes the difference between the predicted RI (y_i') and the ground truth RI y_i. A warm-up strategy has been implemented to increase the stability of the training. Initially, a relatively small learning rate is set and gradually increased to accelerate convergence as the model approaches stability. This warm-up phase is followed by linear decay [46]. For optimization, the Adam algorithm [47] is used to optimize the randomly initialized model

weight parameters on minibatches. Each minibatch is defined as a vector of size $B \times H \times W$, where the batch size is set to 256 and $H \times W$ represents the dimension of the input matrix. In the CNN-based baseline model, the input size is adjusted to $B \times H \times W \times 1$ to meet the dimensional requirements of the convolutional module. Similarly, in the Transformer-Based Baseline model, the input size is transformed to $B \times W \times H$ to accommodate the multi-head attention mechanism.

$$L_{ri} = \begin{cases} 0.5\left((y_i - y_i')^2\right) & |y_i - y_i'| < 1 \\ |y_i - y_i'| - 0.5 & |y_i - y_i'| > 1 \end{cases} \quad (1)$$

4. Evaluation

This section presents the experiments conducted to evaluate the efficacy of baseline models in SARIE. All data used in these experiments were from the proposed SARID. The dataset was partitioned into a training set and a test set at a 7:3 ratio, specifically obtaining 8441 and 3625 audio data samples, respectively. Evaluation and analysis were conducted on the test dataset. The training and evaluation tasks were performed on a GPU server equipped with an Intel® Core™ i7-8700 K C, @3.70 GHz, 64.0 GB of RAM, and an NVIDIA GeForce RTX 2060 G with 6.0 GB of memory.

4.1. Baseline model analysis

For the evaluation metric, we followed the standard protocol in the regression model and used mean absolute error (*MAE*), root mean absolute error (*RMSE*), and coefficient of determination (R^2). These metrics are defined as equations (2)–(4). In the equations, n is the number of samples, y_i and y_i' are the ith ground truth and the corresponding estimated value ($i = 1,2,\ldots,n$), and $\bar{y}$ is the mean of all ground truth data. Lower *MAE* and *RMSE* values and higher R^2 value indicate better prediction performance of the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(|y_i - y_i'|\right) \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i'\right)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i'\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \quad (4)$$

Based on the selection of acoustic feature and network architecture mentioned above, the baseline models could be divided into nine combinations: (a) MFCC-CNN, (b) MFCC-LSTM, (c) MFCC-Transformer, (d) Mel-CNN, (e) Mel-LSTM, (f) Mel-Transformer, (g) STFT-CNN, (h) STFT-LSTM, and (i) STFT-Transformer. Here, the first notation (e.g., MFCC) refers to the input of the network, and the second notation (e.g., CNN) refers to the network. From the results presented in Fig. 6 and Table 3, it can be concluded that MFCC-Transformer outperforms other combinations.

For the three deep learning architectures, the Transformer-based model exhibits significant advantages across various metrics. For example, the model with MFCC inputs achieved an *MAE* of 0.563, *RMSE* of 0.88, and R^2 of 0.765, outperforming CNN and LSTM. Moreover, analysis of the scatter plot (Fig. 6a–c) reveals that the Transformer-based model's fitted line aligns more closely with the ground truth line than the other two architectures. The reason might be that the Transformer can capture more comprehensive features with self-attention than CNN and LSTM, which tend to focus on capturing a single aspect. Benefiting from this attribute, the Transformer-based models can effectively harness the rich structural patterns and intricate relationships present in the audio data.

For the three acoustic features, the performance of the model with the MFCC input surpassed that of the models utilizing Mel and STFT features. For example, in the Transformer-based architecture, compared to Mel features, MFCC exhibited a notable improvement in performance metrics. It demonstrated an enhancement of 16.34% in *MAE*, 16.90% in *RMSE*, and 16.08% in R^2. When compared to the STFT features, MFCC showed a substantial improvement of 20.81% in *MAE*, 19.63% in *RMSE*, and 20.28% in R^2. The MFCC-based model achieved a better fit. In contrast, the STFT-based model tended to underestimate RI, while the Mel-based model tended to overestimate it, particularly at lower RIs (Fig. 6c–f, i). This can be attributed to the sensitivity of Mel filters to low-frequency features, and as RI decreases, the rainfall soundscape often exhibits lower frequency characteristics. However, the performance difference between the two approaches was not significant.

4.2. Baseline model effectiveness analysis

To validate the effectiveness of the baseline model, we conducted a comparative study involving four classical algorithms: decision tree (DT) [48], random forest (RF) [49], linear support vector machine (LSVM) [50], and ridge regression model (RRM) [51]. Using MFCC as the input feature, we selected the Transformer-based baseline model as our benchmark for comparison. The objective was to predict the RI and cumulative rainfall (CR) for three different rainfall events.

The performance of the compared algorithms for estimating RI and CR were provided (Figs. 7 and 8). While these algorithms generally exhibited similar trends, noticeable differences emerged in their estimation accuracy. The baseline model consistently outperformed the other algorithms for RI estimation, demonstrating a closer fit to the ground truth curve and less variation in outliers. Regarding CR estimation, the Transformer baseline model exhibited superior performance compared to alternative regression algorithms, despite a slight overestimation on September 15, 2022. Notably, for the rainfall events on October 5, 2022, and November 21, 2022, the baseline model closely aligned with the reference CR curve, underscoring its effectiveness in capturing cumulative rainfall patterns.

Further insight can be gained by observing the estimation errors of RI and CR (RE_{RI} and RE_{CR}). The *MAE* was used as the metric, which can be calculated as follows:

$$RE_{RI} = \frac{1}{n} \times \left(\sum_{i=1}^{n}\frac{|RI_i' - RI_i|}{RI_i}\right) \times 100\% \quad (5)$$

$$RE_{CR} = \frac{1}{n} \times \left(\sum_{i=1}^{n}\frac{|cr_i' - cr_i|}{cr_i}\right) \times 100\% \quad (6)$$

where RI_i' is the i th reference RI; RI_i is the i th predicted RI calculated by different algorithms; cr_i' is the i th reference CR; cr_i is the i th predicted CR calculated by different algorithms. The results are shown in Tables 4 and 5; the best results are highlighted in bold.

Table 4 indicates the estimation error for different rainfall events for different algorithms. The RE_{RI} for the DT algorithm showed a wide range of errors, ranging from 74.95% to 163.3%. The error of RE_{RI} for the RF algorithm ranged from 25.72% to 42.6%. The

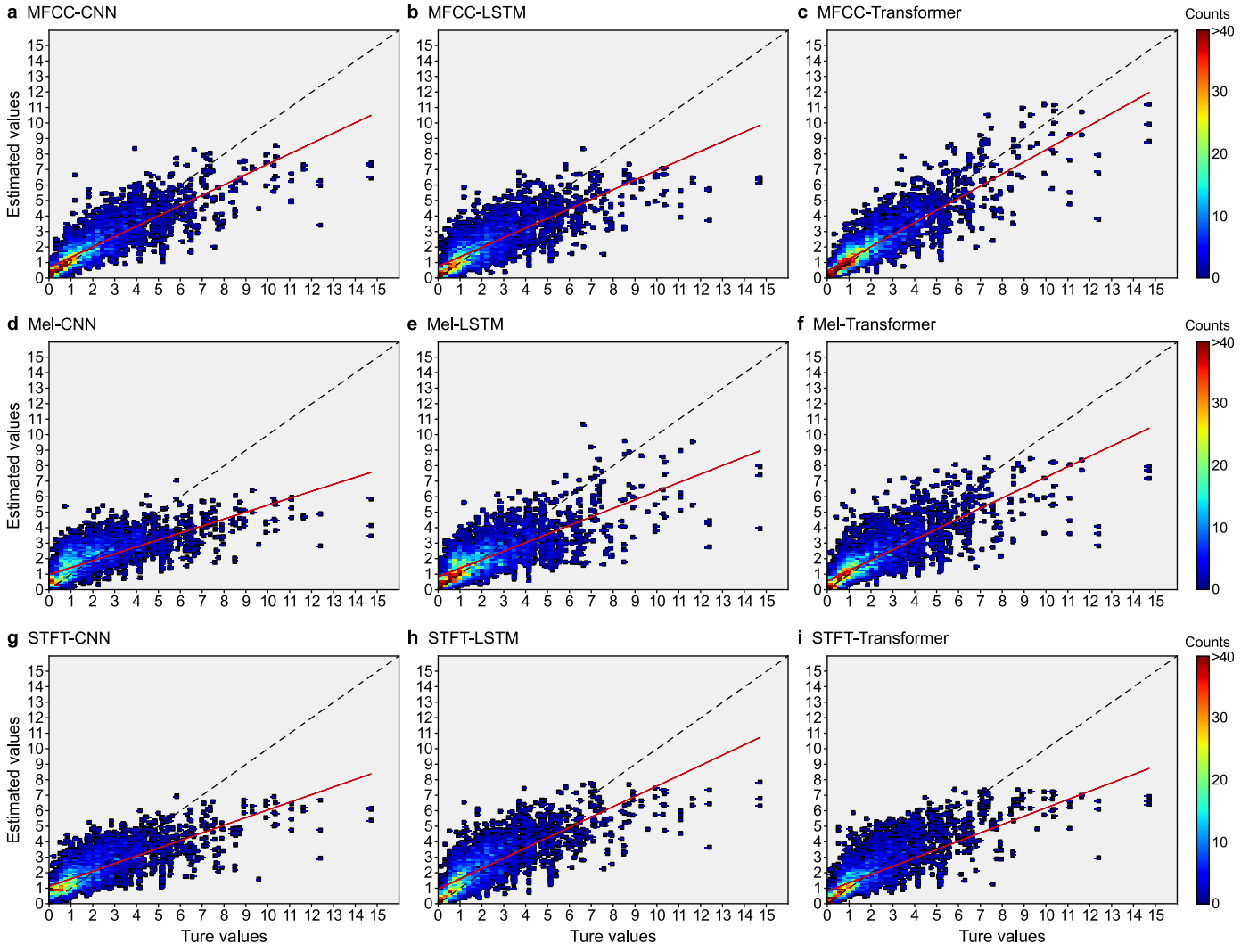

Fig. 6. Rainfall intensity estimation scatterplot of different baseline models.

Table 3
Performance of different baseline models.

Network	Acoustic feature	*MAE* (mm h^{-1})	*RMSE* (mm h^{-1})	R^2
CNN	MFCC	0.646	1.004	0.694
LSTM	MFCC	0.713	1.094	0.637
Transformer	**MFCC**	**0.563**	**0.88**	**0.765**
CNN	Mel	0.85	1.291	0.494
LSTM	Mel	0.796	1.218	0.55
Transformer	Mel	0.673	1.059	0.659
CNN	STFT	0.856	1.223	0.546
LSTM	STFT	0.816	1.163	0.589
Transformer	STFT	0.711	1.095	0.636

LSVM algorithm exhibited RE_{RI} ranging from 20.33% to 111.7%, while the RRM algorithm showed RE_{RI} ranging from 40.00% to 152.18%. In contrast, the RE_{RI} for the baseline model ranged from 15.18% to 33.84%, except for the 15:30 to 17:00 time segment on October 5. Across different rainfall segments, the baseline model consistently outperformed the other algorithms in estimating RI, highlighting its overall superior performance.

Table 5 presents the RE_{CR} in CR estimation. The DT algorithm showed errors ranging from 3.35% to 24.72%, while the RF algorithm showed errors ranging from 1.63% to 16.95%. The LSVM algorithm had errors ranging from 7.1% to 36.88%. In comparison, the error of the baseline model ranged from 11.27% to 21.87%. Across all segments tested, the baseline model consistently outperformed the other regression algorithms in terms of CR error. In summary, the results discussed above underscore the effectiveness of the proposed SARIE baseline models in establishing robust mapping between audio and RI. This demonstrates the potential of SARIE for accurate and high-resolution rainfall estimation.

4.3. Effect of scenario factors on estimation accuracy

All baseline models essentially operate on a global scale, extracting rainfall information from the entirety of acoustic features. The results presented in Sections 4.1 and 4.2 are based on the test dataset without considering scenario factors. Given the complex and diverse nature of urban monitoring of soundscapes, it is crucial to investigate whether these baseline models exhibit biases

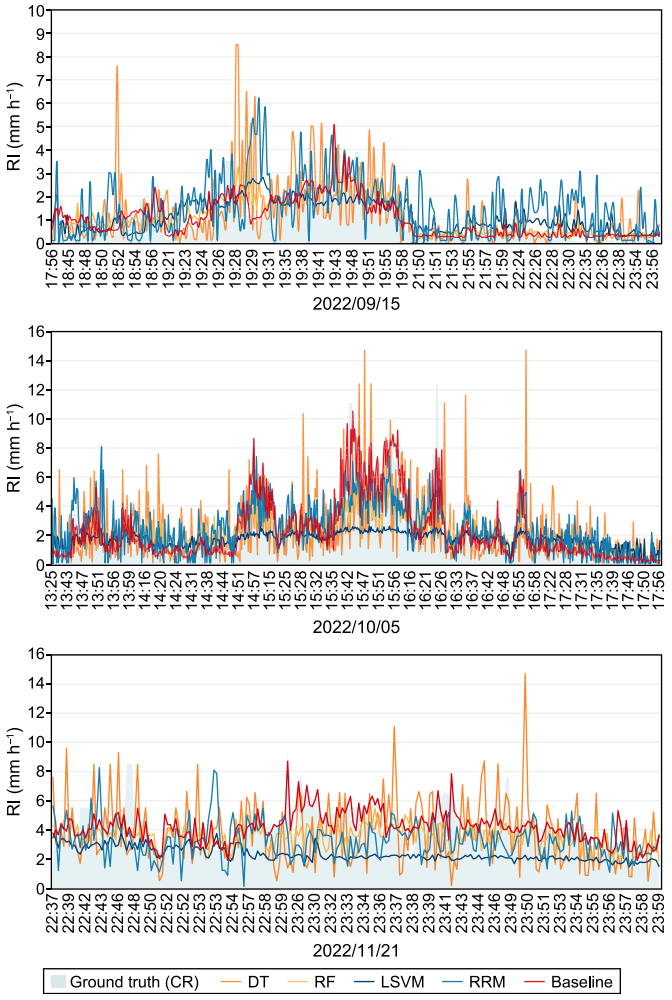

Fig. 7. Experimental results of different algorithms in terms of rainfall intensity.

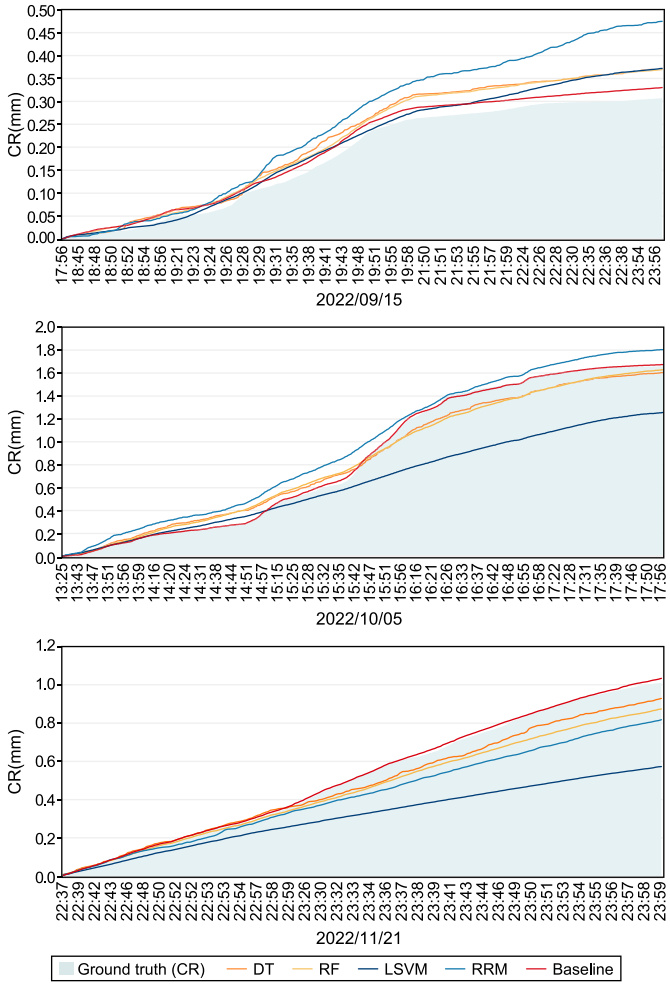

Fig. 8. Experimental results of different algorithms in terms of cumulative rainfall.

in different scenarios. Variation in environmental factors was limited primarily because the recorded rainfall events occurred within the same season, resulting in minimal fluctuations in meteorological conditions (Fig. 4e). Therefore, in this section, we focus on assessing biases toward specific baselines and determining whether the presence of noise affects the accuracy of the results.

We investigated the impact of scenario factors on the performance of the baseline models through the following analysis:

First, we calculated the estimation error values of the baseline models for different underlying surfaces. A box plot was adopted to present the result (Fig. 9). Among the various combinations tested, the MFCC-Transformer combination demonstrated superior accuracy and stability across most underlying surfaces within the baseline models. Notably, models operating on softer surfaces such as “urban meadow” and “water” exhibited better performance. This observation can be attributed to cleaner soundscapes in urban meadow and water environments.

We then evaluated the baseline model with MFCC as the input and Transformer as the network on rainfall audio data, both with and without noise, to assess the impact of different noise conditions. The results were presented (Fig. 10 and Table 6), with all experimental settings matching those in Section 4.1. As depicted, the model trained on data with noise exhibited a 4.9% decrease in

Table 4
Rainfall intensity estimation error (%).

Date	Time	DT	RF	LSVM	RRM	Baseline
09–15	18:45–20:00	114.53	38.66	66.73	152.18	**27.76**
	21:50–22:40	98.12	38.10	55.39	134.18	**33.84**
10–05	15:30–17:00	74.95	25.72	**20.33**	40.82	25.25
	14:55–15:30	163.3	42.6	111.7	40.00	**16.50**
11–21	22:50–23:30	85.89	34.01	76.03	119.81	**15.18**

Table 5
Cumulative rainfall estimation error (%).

Date	Time	DT	RF	LSVM	RRM	Baseline
09–15	18:45–20:00	24.72	21.87	10.63	28.75	**16.95**
	21:50–22:40	15.45	14.94	8.98	26.73	**6.91**
10–05	15:30–17:00	10.14	12.12	36.88	11.27	**1.95**
	14:55–15:30	15.07	17.21	7.1	27.87	**1.63**
11–21	22:50–23:30	3.35	4.9	36.39	18.99	**1.66**

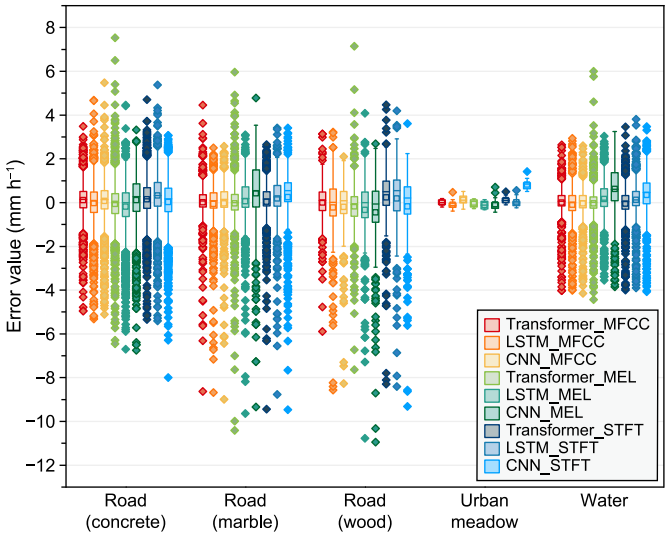

Fig. 9. Error value of baseline models in different underlying surfaces.

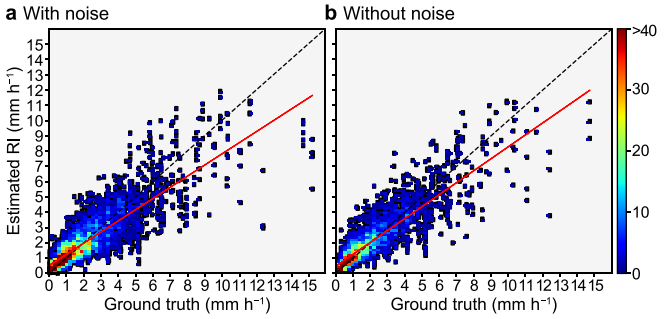

Fig. 10. Rainfall intensity estimation scatterplot of baseline model with/without noise **a**, with noise. **b**, without noise.

R^2, accompanied by increases of 8.6% and 10.1% in *MAE* and *RMSE*, respectively. This suggests that although noise can affect RI estimation, the model's accuracy remains within an acceptable range without significantly reducing precision.

5. Conclusion

In this study, we introduced SARID, a comprehensive and diverse surveillance audio dataset, and an effective baseline model for RI estimation based on surveillance audio. SARID contains meticulously annotated audio samples organized in a consistent format. With a substantial amount of data (12,066 sound slices), diverse data sources (i.e., rainfall, meteorological data, and scenario information), and detailed annotations, SARID represents a valuable resource for advancing research in RI estimation. Our evaluation results demonstrated the effectiveness of the proposed baseline model in mapping surveillance audio to RI. SARID has the potential to open up new avenues for research on rainfall observation applications, particularly in the context of urban surveillance audio. Despite the promising performance of surveillance audio data, our study still has some limitations. First, the analysis of underlying surfaces lacks a cross-sectional analysis of the same rainfall event. Thus, the impact of different underlying surfaces on the surveillance of rainfall audio needs further investigation. Second, some meteorological variables, such as temperature and wind speed, were not considered. Omitting these factors may cause the current model to lack the ability to fully sense the rainfall environment. Third, the current model uses existing acoustic features and does not explore whether there are acoustic features that are more suitable for rainfall observations. Developing an acoustic signature more suitable for rainfall observation is a challenging topic worth further investigation.

Looking ahead, several promising avenues exist for future work with SARID.

(1) Expanding the dataset to include more rainfall audio recordings in different scenarios, especially during extreme rainfall events. A larger dataset will enhance the robustness of the models; we are actively working to collect additional data to further develop SARID.
(2) Improved baseline models: Enhancing the baseline model by incorporating acoustic signals and environmental factors. While we used audio files and basic RI annotations in this study, additional annotations, such as meteorological information, could provide valuable insights to improve rainfall intensity estimation. Moreover, reducing the effect of noise on the model is one of the future research directions.
(3) Integrating surveillance visual and acoustic data into a multimodal approach represents a key strategy. Combining the strengths of both signals offers a promising avenue to enhance rain measurement accuracy. We expect that this multi-model integration will be developed in the future to bring about even stronger performance.

Table 6
Performance of baseline model with/without noise.

Condition	*MAE* (mm h^{-1})	*RMSE* (mm h^{-1})	R^2
With noise	0.616	0.979	0.729
Without noise	0.563	0.88	0.765

Data availability

The source codes and dataset are available for download at the link: https://github.com/Meizhen2023/SARID.

CRediT authorship contribution statement

Meizhen Wang: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing - Original Draft, Writing - Review & Editing. **Mingzheng Chen:** Data Curation, Formal Analysis, Software, Visualization, Writing - Original Draft, Writing - Review & Editing. **Ziran Wang:** Data Curation, Formal Analysis. **Yuxuan Guo:** Data Curation, Formal Cnalysis. **Yong Wu:** Data Curation, Formal Analysis. **Wei Zhao:** Data Curation, Formal Analysis. **Xuejun Liu:** Conceptualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was funded by the National Key R&D Program of China (2021YFE0112300), the State Scholarship Fund from the China Scholarship Council (CSC) (No. 201906865016), and the Special Fund for Public Welfare Scientific Institutions of Fujian Province (No. 2020R1002002).

References

[1] M.N. Anagnostou, E.N. Anagnostou, J.A. Nystuen, S. Michaelides, Application of underwater passive acoustic measurements of ocean sound in precipitation estimation, Precipitation Science, Elsevier, 2022, pp. 37–89.
[2] T. Einfalt, K. Arnbjerg-Nielsen, C. Golz, N.-E. Jensen, M. Quirmbach, G. Vaes, B. Vieux, Towards a roadmap for use of radar rainfall data in urban drainage, J. Hydrol. 299 (3–4) (2004) 186–202.
[3] L. García, D. Rodríguez, M. Wijnen, I. Pakulski, Earth Observation for Water Resources Management: Current Use and Future Opportunities for the Water Sector, World Bank Publications, 2016.
[4] S. Jiang, V. Babovic, Y. Zheng, J. Xiong, Advancing opportunistic sensing in hydrology: a novel approach to measuring rainfall with ordinary surveillance cameras, Water Resour. Res. 55 (4) (2019) 3004–3027.
[5] S. Ochoa-Rodriguez, L.P. Wang, P. Willems, C. Onof, A review of radar-rain gauge data merging methods and their potential for urban hydrological applications, Water Resour. Res. 55 (8) (2019) 6356–6391.
[6] A. Overeem, H. Leijnse, R. Uijlenhoet, Measuring urban rainfall using microwave links from commercial cellular communication networks, Water Resour. Res. 47 (12) (2011).
[7] X. Wang, M. Wang, X. Liu, T. Glade, M. Chen, Y. Xie, H. Yuan, Y. Chen, Rainfall observation using surveillance audio, Appl. Acoust. 186 (2022) 108478.
[8] X. Wang, S. Shi, L. Zhu, Y. Nie, G. Lai, Traditional and novel methods of rainfall observation and measurement: a review, J. Hydrometeorol. 24 (12) (2023) 2153–2176.
[9] G. Bruni, R. Reinoso, N. Van De Giesen, F. Clemens, J. Ten Veldhuis, On the sensitivity of urban hydrodynamic modelling to rainfall spatial and temporal resolution, Hydrol. Earth Syst. Sci. 19 (2) (2015) 691–709.
[10] I. Emmanuel, H. Andrieu, E. Leblois, B. Flahaut, Temporal and spatial variability of rainfall at the urban hydrological scale, Journal of hydrology 430 (2012) 162–172.
[11] B. Shehu, U. Haberlandt, Relevance of merging radar and rainfall gauge data for rainfall nowcasting in urban hydrology, J. Hydrol. 594 (2021) 125931.
[12] M.F. McCabe, M. Rodell, D.E. Alsdorf, D.G. Miralles, R. Uijlenhoet, W. Wagner, A. Lucieer, R. Houborg, N.E. Verhoest, T.E. Franz, The future of Earth observation in hydrology, Hydrol. Earth Syst. Sci. 21 (7) (2017) 3879–3914.
[13] M. Chen, X. Wang, M. Wang, X. Liu, Y. Wu, X. Wang, Estimating rainfall from surveillance audio based on parallel network with multi-scale fusion and attention mechanism, Rem. Sens. 14 (22) (2022) 5750.
[14] J.B. Haurum, C.H. Bahnsen, T.B. Moeslund, Is it raining outside? Detection of rainfall using general-purpose surveillance cameras, in: CVPR Workshops, 2019, pp. 55–64.
[15] R. Zen, D.M.S. Arsa, R. Zhang, N.A.S. Er, S. Bressan, Rainfall estimation from traffic cameras, in: International Conference on Database and Expert Systems Applications, Springer, 2019, pp. 18–32.
[16] R. Avanzato, F. Beritelli, A cnn-based differential image processing approach for rainfall classification, Adv. Sci. Technol. Eng. Syst. J. 5 (4) (2020) 438–444.
[17] X. Wang, M. Wang, X. Liu, L. Zhu, T. Glade, M. Chen, W. Zhao, Y. Xie, A novel quality control model of rainfall estimation with videos – a survey based on multi-surveillance cameras, J. Hydrol. 605 (2022) 127312.
[18] R. Zen, D.M.S. Arsa, R. Zhang, N.A.E.R. Sanjaya, S. Bressan, Rainfall estimation from traffic cameras, in: I. Khalil, Linz (Eds.), Database and Expert Systems Applications, vol. 11706, 2019, pp. 18–32. AUSTRIA.
[19] P. Allamano, A. Croci, F. Laio, Toward the camera rain gauge, Water Resour. Res. 51 (3) (2015) 1744–1757.
[20] K. Garg, S.K. Nayar, Vision and rain, Int. J. Comput. Vis. 75 (1) (2007) 3–27.
[21] K. Garg, S.K. Nayar, When does a camera see rain?, in: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, vol. 2 IEEE, 2005, pp. 1067–1074.
[22] H. Yin, F. Zheng, H.-F. Duan, D. Savic, Z. Kapelan, Estimating rainfall intensity using an image-based deep learning model, Engineering 21 (2023) 162–174.
[23] F. Zheng, H. Yin, Y. Ma, H.F. Duan, H. Gupta, D. Savic, Z. Kapelan, Toward improved real-time rainfall intensity estimation using video surveillance cameras, Water Resour. Res. 59 (8) (2023) e2023WR034831.
[24] J.A. Nystuen, Underwater Ambient Noise Measurements of Rainfal, University of California, San Diego: United States – California, 1985, p. 98.
[25] J.A. Nystuen, Rainfall measurements using underwater ambient noise, J. Acoust. Soc. Am. 79 (4) (1986) 972–982.
[26] B.B. Ma, B.D. Dushaw, B.M. Howe, Rainfall at sea: using the underwater sounds of raindrops as a rain gauge for weather and climate, Acoust. Today (2022) 62–71.
[27] D. Dunkerley, Acquiring unbiased rainfall duration and intensity data from tipping-bucket rain gauges: a new approach using synchronised acoustic recordings, Atmos. Res. 244 (2020) 105055.
[28] R. Avanzato, F. Beritelli, A. Raspanti, M. Russo, Assessment of multimodal rainfall classification systems based on an audio/video dataset, Int. J. Adv. Sci. Eng. Inf. Technol. 10 (2020) 1163–1168.
[29] X. Wang, T. Glade, E. Schmaltz, X. Liu, Surveillance audio-based rainfall observation: an enhanced strategy for extreme rainfall observation, Appl. Acoust. 211 (2023) 109581.
[30] E.M. Trono, M.L. Guico, N.J.C. Libatique, G.L. Tangonan, D.N.B. Baluyot, T.K.R. Cordero, F.A.P. Geronimo, A.P.F. Parrenas, Rainfall monitoring using acoustic sensors, in: TENCON 2012 IEEE Region 10 Conference, 2012, pp. 1–6.
[31] C. Bedoya, C. Isaza, J.M. Daza, J.D. López, Automatic identification of rainfall in acoustic recordings, Ecol. Indicat. 75 (2017) 95–100.
[32] M.L. Guico, G. Abrajano, P.A. Domer, J.P. Talusan, Design and Development of a Novel Acoustic Rain Sensor with Automated Telemetry, 2018.
[33] O.C. Metcalf, A.C. Lees, J. Barlow, S.J. Marsden, C. Devenish hardRain, An R package for quick, automated rainfall detection in ecoacoustic datasets using a threshold-based approach, Ecol. Indicat. 109 (2020) 105793.
[34] C. Sánchez-Giraldo, C.L. Bedoya, R.A. Morán-Vásquez, C.V. Isaza, J.M. Daza, Ecoacoustics in the rain: understanding acoustic indices under the most common geophonic source in tropical rainforests, Remote Sensing in Ecology and Conservation 6 (3) (2020) 248–261.
[35] M. Ferroudj, A. Truskinger, M. Towsey, L. Zhang, J. Zhang, P. Roe, Detection of rain in acoustic recordings of the environment, in: D.-N. Pham, S.-B. Park (Eds.), Pacific Rim International Conference on Artificial Intelligence 2014: Trends in Artificial Intelligence, Springer International Publishing: Gold Coast, Queensland, Australia, 2014, pp. 104–116.
[36] A. Brown, S. Garg, J. Montgomery, Automatic rain and cicada chorus filtering of bird acoustic data, Appl. Soft Comput. 81 (2019) 105501.
[37] R. Avanzato, F. Beritelli, An innovative acoustic rain gauge based on convolutional neural networks, Information 11 (2020).
[38] R. Avanzato, F. Beritelli, F.D. Franco, V.F. Puglisi, A convolutional neural networks approach to audio classification for rainfall estimation, in: 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, vol. 1, IDAACS, Metz, France, 2019, pp. 285–289.
[39] R. Avanzato, F. Beritelli, A. Raspanti, M. Russo, Assessment of multimodal rainfall classification systems based on an audio/video dataset, Int. J. Adv. Sci. Eng. Inf. Technol. 10 (2020) 1163.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems vol. 30, 2017, pp. 6000–6010.
[41] P. Liptai, M. Badida, K. Lukáčová, Influence of atmospheric conditions on sound propagation-mathematical modeling, Óbuda University e-Bulletin 5 (1) (2015) 127.
[42] S. Chu, S. Narayanan, C.-C.J. Kuo, Environmental sound recognition with time–frequency audio features, IEEE Trans. Audio Speech Lang. Process. 17 (6) (2009) 1142–1158.
[43] Z.K. Abdul, A.K. Al-Talabani, Mel frequency cepstral coefficient and its applications: a review, IEEE Access 10 (2022) 122136–122158.
[44] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: audio and music signal analysis in python, in: Proceedings of the 14th python in Science Conference, vol. 8, 2015, pp. 18–25.
[45] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[46] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, Accurate, Large Minibatch Sgd: Training Imagenet in 1 Hour, 2017 arXiv preprint arXiv:1706.02677.
[47] D.P. Kingma, J. Ba Adam, A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.
[48] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, An introduction to decision tree modeling, J. Chemometr.: A Journal of the Chemometrics Society 18 (6) (2004) 275–285.
[49] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2016) 197–227.
[50] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (2) (1998) 121–167.
[51] A.E. Hoerl, R.W. Kennard, Ridge regression: applications to nonorthogonal problems, Technometrics 12 (1) (1970) 69–82.