International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 3
ISSN: 2088-5334

Assessment of Multimodal Rainfall Classification Systems Based on an Audio/Video Dataset

Roberta Avanzato[a,1], Francesco Beritelli[a,2], Antonio Raspanti[b,3], Michele Russo[b,4]

[a] *Department of Electrical, Electronic and Informatics Engineering, University of Catania, Catania, 95125, Italy*
E-mail: [1]roberta.avanzato94@hotmail.it; [2]francesco.beritelli@dieei.unict.it

[b] *VICOSYSTEMS S.R.L., Viale Odorico da Pordenone, Catania, 95128, Italy*
E-mail: [3]a.raspanti@vicosystems.it; [4]m.russo@vicosystems.it

Abstract— **In the past few years, there has been an increase in natural disasters due to hydrogeological instability caused by heavy rain. Therefore, to reduce the risk of an imminent occurrence of a disastrous event and reduce the risk to humans, an accurate estimate of the precipitation levels based on advanced machine learning techniques is necessary. In this paper, a new dataset is proposed containing audio/video data recorded via a multimodal rain gauge created ad hoc. The dataset, denominated AVDB-4RC (Audio/Video Database for Rainfall Classification), contains digital audio/video sequences recorded for seven different levels of precipitation intensity. In particular, the database presents a set of audio sequences containing the acoustic timbre produced by the rain and video sequences containing rain videos, both in seven different intensities, i.e., "No rain," "Weak rain," "Moderate rain," "Heavy rain" and "Very heavy rain," "Shower rain" and "Cloudburst rain." For the validation of the dataset, the paper proposes a novel rainfall classification approach based on a video pattern recognition system that uses CNN neural networks. The average classification accuracy is approximately 49% and can reach 75% if the adjacent misclassifications are not considered. Presumably, it is the first open dataset from the new generation acoustic/video rain gauges available for evaluating the estimated rainfall performance. We hope that this new open dataset will encourage a comparison of rainfall estimation/classification algorithms on this common database so that the adopted techniques are objectively assessed and improved.**

Keywords— **audio/video database; rainfall classification; open dataset; performance evaluation; convolutional neural network.**

I. INTRODUCTION

The creation of multimodal datasets (e.g., audio, video) containing information regarding the levels of rainfall intensity is essential to ensure the safety of people and things in hydrogeological risk management scenarios. Thus, innovative and accurate techniques of rainfall classification must be employed. Tilt rain gauges generally consist of a plastic manifold balanced on a pin. When it tips, it actuates a switch; this action is then electronically recorded or transmitted to a remote collection station. The disadvantage of this system lies in the fact that it tends to underestimate the amount of rainfall, particularly in snowfall and heavy rainfall events. Besides, the inclination of the receiver and different types of dirt that may clog the water collection point also has an effect on its performance.

These problems have led to the use of alternative rain gauges capable of facing this challenge, such as weather radar, satellite, and radio link rain gauge [1]–[4].

To overcome the limitations present in traditional techniques of classification, recent studies [5]–[9] have adopted advanced neural network approaches and audio identification features [10]–[12]. Further studies have devised rain gauges based on the analysis of rain images to classify rainfall intensity [13]–[16].

This paper presents a new database that includes audio and video sequences and the first example of a Convolutional Neural Network (CNN) approach to the classification of rainfall levels through differential video images. The new dataset has been denominated Audio/Video Database for Rainfall Classification (AVDB-4RC), and it contains data collected from seven different levels of rainfall that can be used for designing and evaluating the performance of multimodal rain gauge systems.

In particular, the database presents a set of audio sequences containing the acoustic timbre produced by the rain and video sequences containing rain videos, both in seven different intensities, i.e., "No rain," "Weak rain," "Moderate rain," "Heavy rain" and "Very heavy rain," "Shower rain" and "Cloudburst rain." The database proposed in this paper is an extension of the one used in our previous study [17], where we proposed a new rainfall classification algorithm based on an audio signal and evaluated its performance. At first, the spectral and statistical analysis of the audio sequences were presented, followed by the performances from the accuracy of an audio pattern recognition system based on the CNN network.

In this paper, we carry out a similar study focusing on the recorded video sequences. At first, we define the structure of the proposed dataset, and then we evaluate the database containing the video sequences by training a convolutional neural network. The paper is organized as follows: Section II illustrates the rainfall classification audio and video dataset; Section III describes the hardware and software components used; Section IV illustrates the audio and video dataset before feeding it to the neural network; Section V shows the CNN adopted for the video dataset in input; Section VI shows the results of the video sequence classification; Section VII depicts the structure of the audio/video database; Section VIII is devoted to conclusions.

II. Material and Method

A. *Rainfall Classification Audio and Video Dataset*

Fig. 1 shows the acquisition system used for the created database. The sequences so labeled are archived in the database. The multimodal acquisition system is divided into two parts:

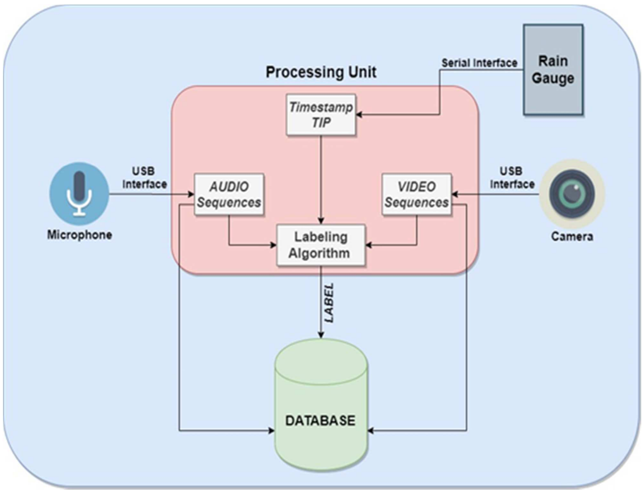


Fig. 1 Audio and Video acquisition system

1) *Audio acquisition system:* the system samples audio sequences at a frequency of 22.05 kHz at 16 bit (PCM format). The database consists of seven precipitation intensity categories, defined in Table 1; for each category, there are 2 audio sequences, lasting 22 seconds each.

2) *Video acquisition system*: the system samples at a fixed frame-rate of 15 FPS with a frame size of 640×480 pixels. The database is made up of seven categories of rainfall intensity, defined in Table 1 for each category 2 video sequences are lasting 22 seconds each, synchronized with the relative audio sequences described above. The classification levels are obtained based on different levels of precipitation. These levels are obtained by taking the national classification scales [18] as a reference and slightly modifying some sample ranges.

In this way, we obtain some adequate and homogeneous samples corresponding to different precipitation classes. The current database was created by recording the sequences on a rainy day characterized by all seven levels of rainfall intensity. Furthermore, during the signal acquisition and recording phase, continuous checks were made on the correct functioning of the rain gauge (e.g., presence of dirt in the collection tray) used to measure and label the rainfall intensity levels.

TABLE I
Rainfall Intensity Classification

Rain Classification	**Precipitation Intensity**
NR – No Rain	< 0.5 mm/h
W – Weak rain	[0.5 – 2] mm/h
M – Moderate rain	[2 – 6] mm/h
H – Heavy rain	[6 – 10] mm/h
VH – Very heavy rain	[10 – 18] mm/h
S – Shower rain	[18 – 30] mm/h
C – Cloudburst rain	> 30 mm/h

Finally, the database was verified and processed manually, eliminating incorrectly labeled sequences, through a repeated phrase of audio/video analysis by an operator.

The dataset is composed of:

- Ten audio-video files for the categories "No rain," "Weak rain," "Moderate rain," "Heavy rain" and "Very heavy rain";
- 3 audio-video files for the "Shower" category;
- 2 audio-video files for the "Cloudburst" category.

70% of the audio and video files make up the learning dataset, the remaining 30% make up the testing dataset. Consequently, the learning dataset is divided into training dataset (70%) and validation dataset (30%).

Once the dataset is created, the audio and video signals are normalized using mean and standard deviation and fed as input to the neural network. The percentage of probability corresponding to each individual class will appear in the output. The fact that the presented database contains video sequences recorded in a single location does not represent a critical issue as it is always possible to make the database independent of the specific background image by conducting differential analysis of the video signals before they are fed as input to the neural network.

The image processing procedures carried out on the videos are as follows:

- Extraction of frames with frame-rate equal to 30 FPS and offset equal to 1 frame;
- Calculation of "differential images".

Fig. 2 shows an example of a differential image for each level of rainfall. Note that no information appears in the background and the number of drops is proportional to the rainfall intensity. DCT is applied to the entire image as a result of the procedure described in the previous points. The data obtained from the application of the DCT are standardized and fed to the neural network.

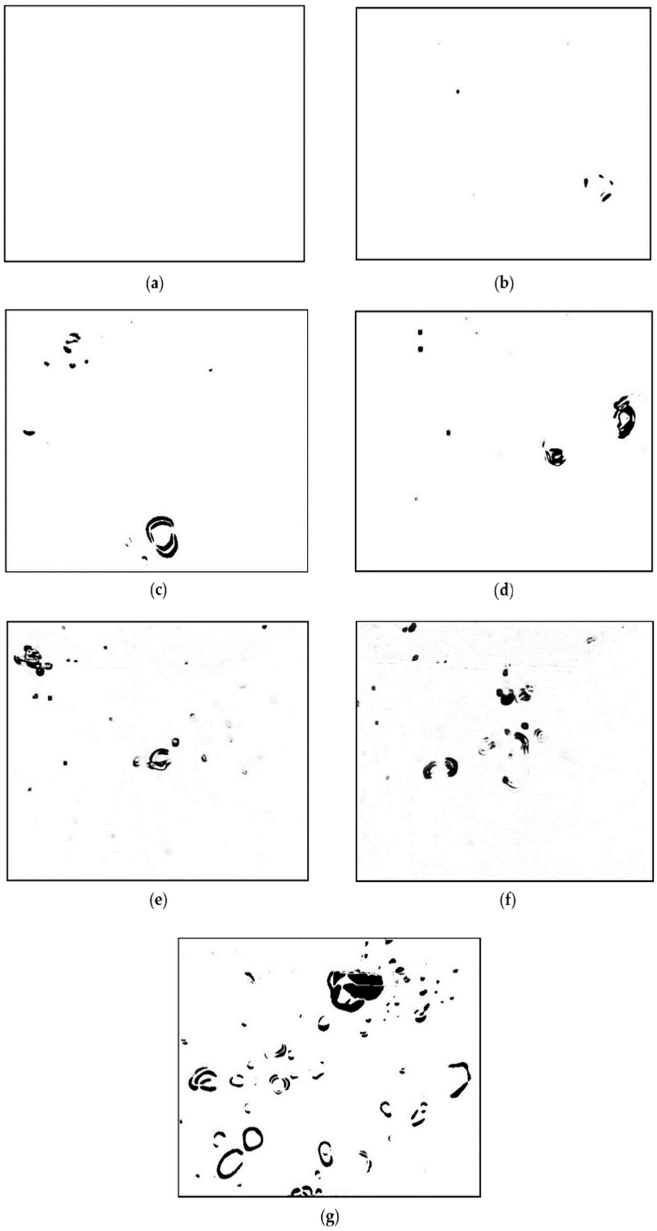

(a) (b) (c) (d) (e) (f) (g)

Fig. 2 Examples of differential frames related to each level of precipitation: (a) “No rain”; (b) “Weak rain”; (c) “Moderate rain”; (d) “Heavy rain”; (e) “Very heavy rain”; (f) “Shower”; (g) “Cloudburst”.

B. *Hardware and Software Description*

The proposed system is capable of detecting the audio and video data of the rain using a video camera with a microphone located inside a plastic shaker. The main interest lies in capturing audio and image data when rain falls on the plastic surface of the shaker. Moreover, we are interested in obtaining a more efficient and rapid classification of the different levels of rainfall intensity. Fig. 3 shows the general scheme of our audio and video sequence acquisition system.

In particular, the system is made up of a microphone and camera (a); plastic shaker with transparent cover (b); tipping bucket rain gauge (c); Raspberry Pi (d), used for data processing; 4G dongle (e) for data transmission.

The dimensions of the hardware components and the distances used by the recording kit are as follows:

- size of the base of the kit equal to 30x40 cm;
- the lower base diameter of the plastic shaker equal to 7 cm;
- the upper base diameter of the plastic shaker equal to 10 cm;
- height of the shaker equal to 17 cm;
- transparent plastic dome thickness equal to 2 mm;
- the distance of the camera/microphone from the sidewall of the shaker equal to 5 cm;
- the diameter of the transparent dome equal to 12 cm;
- the distance of the camera/microphone from the transparent dome equal to 5 cm.

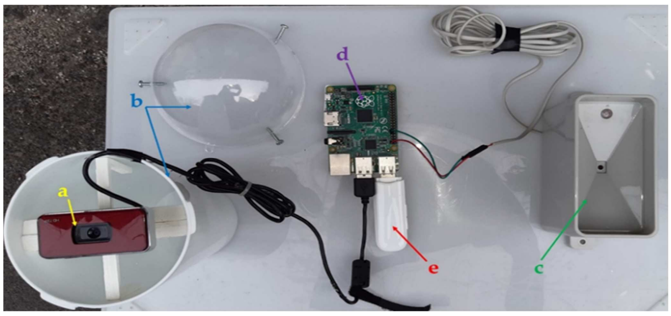


Fig. 3 Hardware components

The kit was placed on a wall raised from the ground at the height of about 1.80 meters and away from sources of disturbance (roads, people, etc.). A camera and a microphone connected to a processing unit make up the system. The processing unit allows the synchronous aggregation of audio and video sequences in the same precipitation intensity class. The labeling algorithm, described in detail in our previous paper [17], allows defining the precipitation intensity classes to which the audio and video files (each 22 seconds long) belong. The labeling takes place utilizing the "interruptions" generated by the tipping bucket type rain gauge, whenever it is overturned by rain.

The tipping rain gauge is connected to the processing unit via an RJ11 cable and is managed ad hoc through a software interface capable of detecting and counting the "interruptions." Once the data is obtained, it can be sent via the cloud (4G dongle) to the cloud and entered into the database [19]. Subsequently, once the rain gauge recordings and "interruptions" have been obtained, it is possible to estimate the quantity of mm/h using formula 1, where the $\frac{T_h}{T_s}$ Ratio corresponds to the number of tips that would occur in an hour-long observation interval assuming constant rain. So, every minute, the amount of rainfall in mm/h relative to that minute, i.e., the 22 seconds of the audio-video sequence contained in that minute, is estimated and labeled in the database according to the rainfall classes defined in Table 1.

$$e = \frac{C{*}T_h}{T_s} \qquad (1)$$

- C [mm] is the capacity of the tipping bucket rain gauge for water collection;
- R_k stands for the instant of the k-th bucket rotation, subsequent to T_i (final instant of the i-th time window of the audio-video signal to be labeled);

- R_{k-1} is the instant of the k-th -1 bucket rotation, prior to T_i;
- T_s corresponds to the time interval between R_k and R_{k-1};
- T_h is 3600 seconds.

This process is repeated for each 22-second recording sequence stored in the database. Every minute the obtained values are sent to an IoT platform via the publish/subscribe protocol [20] and are used to label the audio/video signal at different rainfall intensities.

C. Dataset Structure

The dataset is organized into two folders: Learning and Testing, as shown in Figure 4. The internal structure of the two folders is the same, i.e., both of them contain 7 subfolders named using the initials of each level of precipitation (nr, w, m, h, vh, s, and c). Within each subfolder, relating to each level of precipitation, there are two files, one audio (.wav) and one video (.mkv), both lasting 22 seconds.

D. Audio Dataset

The audio sequences recorded and acquired by the acquisition system described in Section II, are 22 seconds long each. Before being fed as input to the neural network, they are divided into subsequences of shorter duration and a sliding window with an offset of 100 milliseconds is applied. Subsequently, the obtained sub-sequences are fed as input to the neural network. The result of this training and subsequent testing is described in detail in [17].

E. Video Dataset

Before being fed to the CNN, the video sequences undergo a preprocessing phase. As described in Section II, the differential frame sequences are extrapolated, taken at 30 FPS frame-rate, with 1 frame offset. Subsequently, each grayscale frame is initially scaled non-linearly to the size of 224x224 (to correspond to the shape required for the CNN input) and then converted into a matrix. To extract the frequency content of each frame, the DCT (Discrete Cosine Transform) is applied to the matrix.

Finally, certain steps are taken to insert standardized matrices into the neural networks. The first step is to extract frames at a frame rate of 30 FPS from the original video. Next, a difference between the extracted frames is created moving from frame to frame with an offset of 1 frame. The obtained differential images are resized to a size of 224x224.

The image is represented as an array by applying the "NUMPY" function in python. Subsequently, the DCT is applied to the values represented in the matrix, whereas the obtained values are standardized and fed as input to the CNN network. Once the manipulation is complete, a division is performed before feeding all the data to the neural network: 70% of these matrices are inserted in the learning set (further broken down into 70% of the training set and 30% of the validation set) and 30% in the testing set.

Once the dataset is created, each matrix is fed as input to the neural network (CNN). The probability percentage corresponding to each individual class will appear in the output. Section VI will analyze the performance of this network when this validation dataset is applied to the network input.

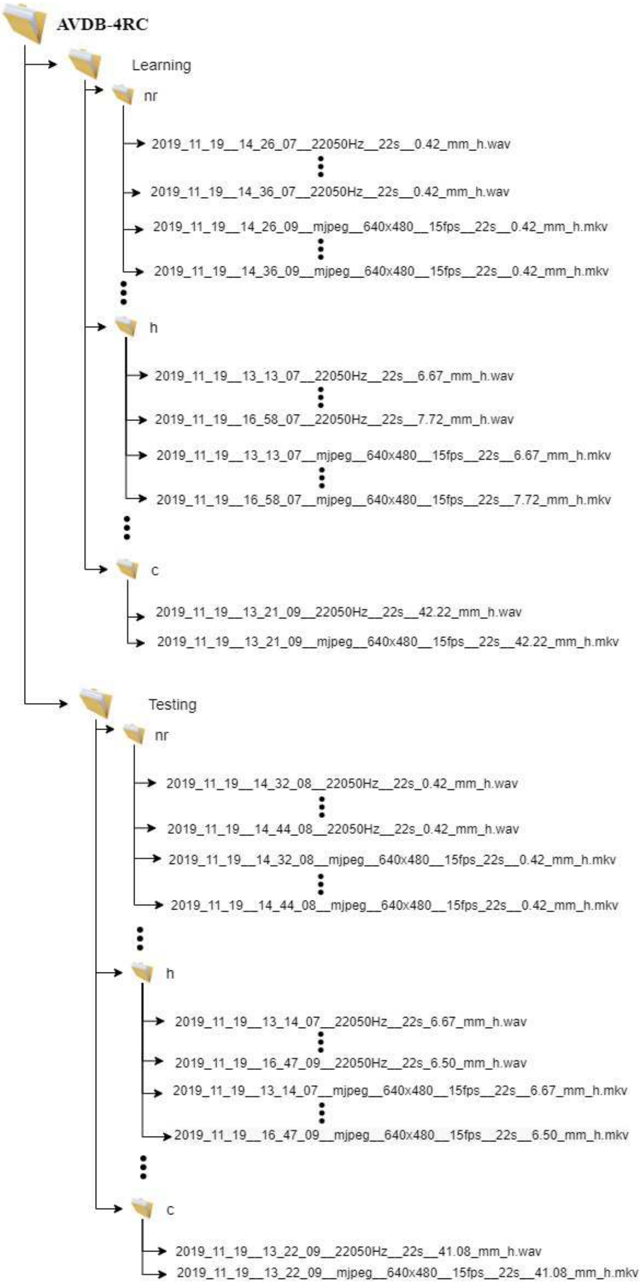

AVDB-4RC
Learning
nr
2019_11_19__14_26_07__22050Hz__22s__0.42_mm_h.wav
2019_11_19__14_36_07__22050Hz__22s__0.42_mm_h.wav
2019_11_19__14_26_09__mjpeg__640x480__15fps__22s__0.42_mm_h.mkv
2019_11_19__14_36_09__mjpeg__640x480__15fps__22s__0.42_mm_h.mkv
h
2019_11_19__13_13_07__22050Hz__22s__6.67_mm_h.wav
2019_11_19__16_58_07__22050Hz__22s__7.72_mm_h.wav
2019_11_19__13_13_07__mjpeg__640x480__15fps__22s__6.67_mm_h.mkv
2019_11_19__16_58_07__mjpeg__640x480__15fps__22s__7.72_mm_h.mkv
c
2019_11_19__13_21_09__22050Hz__22s__42.22_mm_h.wav
2019_11_19__13_21_09__mjpeg__640x480__15fps__22s__42.22_mm_h.mkv
Testing
nr
2019_11_19__14_32_08__22050Hz__22s_0.42_mm_h.wav
2019_11_19__14_44_08__22050Hz__22s_0.42_mm_h.wav
2019_11_19__14_32_08__mjpeg__640x480__15fps_22s__0.42_mm_h.mkv
2019_11_19__14_44_08__mjpeg__640x480__15fps_22s__0.42_mm_h.mkv
h
2019_11_19__13_14_07__22050Hz__22s_6.67_mm_h.wav
2019_11_19__16_47_09__22050Hz__22s_6.50_mm_h.wav
2019_11_19__13_14_07__mjpeg__640x480__15fps__22s__6.67_mm_h.mkv
2019_11_19__16_47_09__mjpeg__640x480__15fps__22s__6.50_mm_h.mkv
c
2019_11_19__13_22_09__22050Hz__22s__41.08_mm_h.wav
2019_11_19__13_22_09__mjpeg__640x480__15fps__22s__41.08_mm_h.mkv

Fig. 4 Learning and testing dataset folder

F. CNN Adopted

The type of convolutional neural network used in this article is a convolutional network already known in state of the art: SqueezeNet [21], which is the evolution of the AlexNet network [22]. The difference between the two is given by the reduced complexity of SqueezeNet compared to AlexNet. In fact, the goal of SqueezeNet is to deliver the same performance of AlexNet using 50 parameters less, thus employing only 5 MB of parameters, instead of 240 MB used by AlexNet. Squeezenet is not an AlexNet compression, rather it is a completely different Deep Neural Network (DNN) architecture.

Both networks have a common denominator, i.e., the level of classification accuracy on the same database (Imagenet).

SqueezeNet is a completely convolutional neural network with reduced complexity and with layers of dropout, which allows improving performance without affecting accuracy. In Fig. 5 it is possible to view the SqueezeNet architecture applied. The first layer is a standalone convolution layer (conv1), followed by 8 Fire modules (fire2-9); the last layer is a convolutional layer (conv10).

The number of filters per fire module from the beginning to the end of the network gradually increases from the first to the last layer. SqueezeNet performs max pooling with a stride of 2 after layers conv1, fire4, fire8, and conv10. Details about this deep neural network are illustrated in [21], [23].

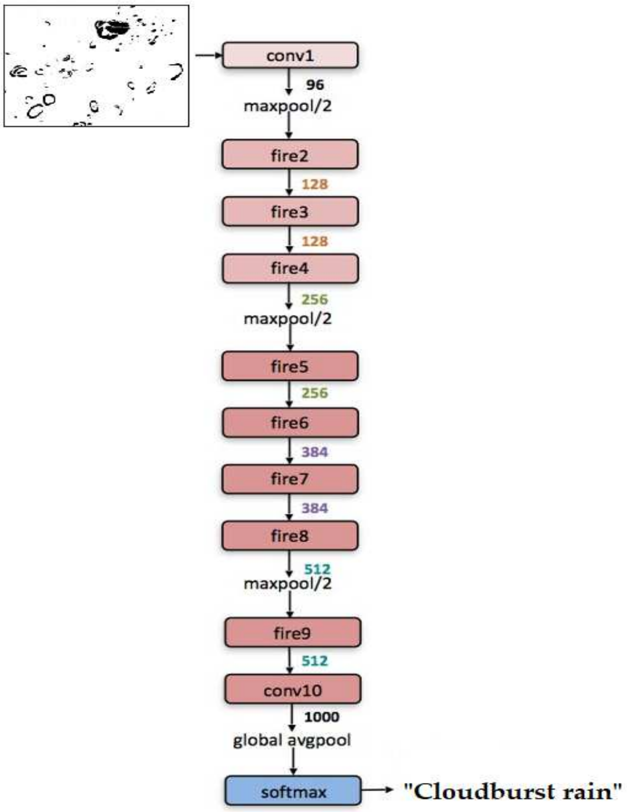


Fig. 5 Microarchitectural view of our SqueezeNet architecture.

III. Results and Discussion

In this section, the results obtained by applying the automatic learning technique will be analyzed, with pre-processed video frames, coming from the validation dataset, fed as input for the CNN. Fig. 6 shows the progress of training losses (blue curve) and test losses (orange curve); both curves decrease with increasing epochs.

On the contrary, Fig. 7 shows that the trend of training accuracy and test accuracy increases with increasing epochs. The two graphs are complementary, since the accuracy increases, the loss for training and tests decreases. This implies that the neural network is performing an accurate classification. Fig. 8 shows the confusion matrix obtained by processing the sequences contained in the validation dataset and entered in the CNN network. The confusion matrix allows calculating the percentage of classification accuracy. It shows that the average classification accuracy is approximately 49% and can reach 75% if the adjacent misclassifications are not considered.

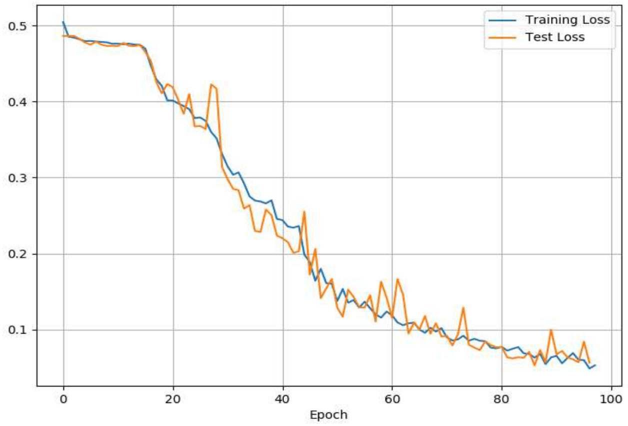


Fig. 6 Training and test losses without sub-block 16x16.

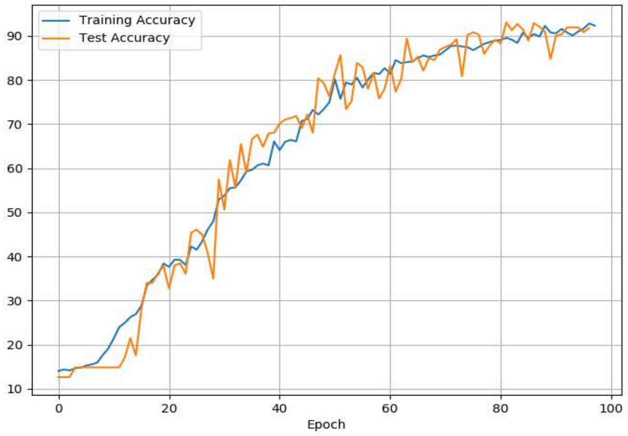
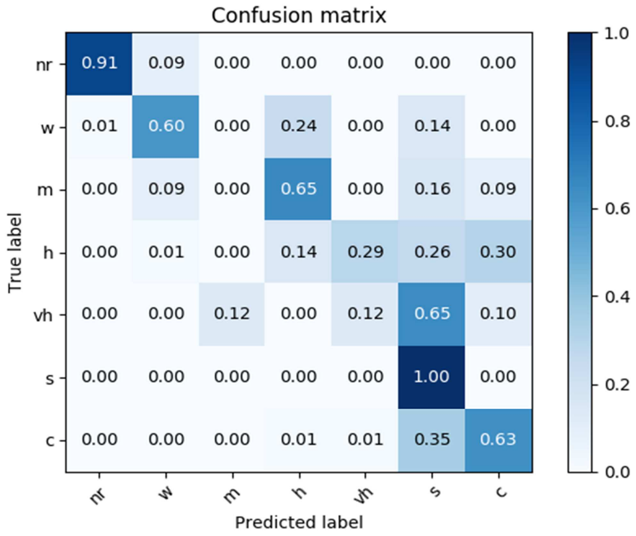


Fig. 7 Training and test accuracy without sub-block 16x16.



	nr	w	m	h	vh	s	c
nr	0.91	0.09	0.00	0.00	0.00	0.00	0.00
w	0.01	0.60	0.00	0.24	0.00	0.14	0.00
m	0.00	0.09	0.00	0.65	0.00	0.16	0.09
h	0.00	0.01	0.00	0.14	0.29	0.26	0.30
vh	0.00	0.00	0.12	0.00	0.12	0.65	0.10
s	0.00	0.00	0.00	0.00	0.00	1.00	0.00
c	0.00	0.00	0.00	0.01	0.01	0.35	0.63

Fig. 8 Confusion matrix without sub-block 16.

IV. Conclusion

The paper presents a freely available set of data for the design and/or validation of innovative mono/multimodal systems for rainfall classification. To the best of our knowledge, this is the only dataset of this kind. Presumably, it is the first open dataset from the new generation acoustic/video rain gauges available for evaluating the estimated rainfall performance. We hope that this new open dataset will encourage a comparison of rainfall estimation/classification algorithms on this common database so that the adopted techniques are objectively assessed and improved. We finally hope that other research groups will contribute to future releases of AVDB-4RC or will release their datasets to the research community for future systems to be tested on open corpora of rainfall sounds/video, increasing the validity of each new scientific result obtained. The audio-video dataset is freely downloadable from the following link: https://github.com/vicosystems/AVDB-4RC.

References

[1] J.M. Trabal, and D.J. McLaughlin, "Rainfall Estimation and Rain Gauge Comparison For X-Band Polarimetric CASA Radars," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2017, pp. 2726-2729.

[2] D. Nagel, "Detection of Rain Areas with Airborne Radar," in *Proc. 18th International Radar Symposium (IRS)*, 2017, pp. 1-7.

[3] A. K. Shukla, C.S.P. Ojha, and R. D. Garg, "Comparative Study ff TRMM Satellite Predicted Rainfall Data with Rain Gauge Data Over Himalayan Basin," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGRSS)*, 2018, pp. 9347-9350.

[4] A. K. Varma, "Measurement of Precipitation from Satellite Radiometers (Visible, Infrared, and Microwave): Physical Basis, Methods, and Limitations," *Remote Sensing of Aerosols, Clouds, and Precipitation*, pp. 223–248, 2018.

[5] F. Beritelli, G. Capizzi, G. Lo Sciuto, F. Scaglione, D. Połap, and M. Woźniak, "A Neural Network Pattern Recognition Approach to Automatic Rainfall Classification by Using Signal Strength in LTE/4G Networks," in *Proc. International Joint Conference on Rough Sets*, 2017, pp. 505-512.

[6] F. Beritelli, G. Capizzi, G. Lo Sciuto, C. Napoli, and F. Scaglione, "Rainfall estimation based on the intensity of the received signal in a LTE/4G mobile terminal by using a probabilistic neural network," *IEEE Access*, vol. 6, pp. 30865–30873, May 2018.

[7] R. Avanzato, F. Beritelli, F. Di Franco, and V.F. Puglisi, "A Convolutional Neural Networks approach to Audio Classification for Rainfall Estimation," in *Proc. 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, 2019, pp. 285-289.

[8] A. Gupta, A. Bansal, R. Gupta, D. Naryani, and A. Sood, "Urban Waterlogging Detection and Severity Prediction Using Artificial Neural Networks," in *Proc. IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2017, pp. 42-49.

[9] A. H. Manek, and P.K. Singh, "Comparative Study of Neural Network Architectures for Rainfall Prediction," in *Proc. IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, 2016, pp. 171-174.

[10] F. Beritelli, and A. Spadaccini, "Statistical Approach to Biometric Identity Verification based on Heart Sounds," *in Proc. Fourth International Conference on Emerging Security Information, Systems and Technologies*, 2010, pp. 93-96.

[11] F. Beritelli, and A. Spadaccini, "The Role of Voice Activity Detection in Forensic Speaker Verification," in *Proc. 17th IEEE International Conference on Digital Signal Processing (DSP)*, 2011.

[12] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very Deep Convolutional Neural Network for Raw Waveforms," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2017.

[13] S. Sawant, and P. A. Ghonge, "Estimation of rain drop analysis using image processing, electronics and telecommunication," *International Journal of Science and Research (IJSR)*, vol. 4, pp. 1981–1986, Jan. 2015.

[14] P. Bacche, S. Basantani, P. Deshpande, J. Joshi, and R. Bhalwankar, "Measurement of raindrop parameters using image processing," *International Journal of Innovative and Emerging Research in Engineering*, vol. 3, pp. 41–46, 2016.

[15] K. H. V. Reddy, S.M. Basha, and J. Srinivasulu, "A simple approach for efficient detection and estimation of drops during the rainfall," *IJISET - International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 10, pp. 203–207, Oct. 2015.

[16] F. Nashashibi, R. De Charette, and A. Lia, "Detection of Unfocused Raindrops on a Windscreen using Low Level Image Processing," in *Proc.11th International Conference on Control Automation Robotics & Vision*, 2010.

[17] R. Avanzato, and F. Beritelli, "An Innovative Acoustic Rain Gauge Based on Convolutional Neural Networks," *MDPI Information*, vol. 11, no. 4, pp. 183, March 2020.

[18] (2020) Smaniotto. [Online]. Available. http://www.smaniotto.eu/scale-della-natura.html.

[19] F. Beritelli, A. Gallotta, and C. Rametta "A Dual Streaming Approach for Speech Quality Enhancement of VoIP Service Over 3G Networks," in *Proc. IEEE International Conference on Digital Signal Processing (DSP)*, 2013.

[20] C. Chen, Y. Tock, and S. Girdzijauskas, "BeaConvey: Co-Design of Overlay and Routing for Topic-based Publish/Subscribe on Small-World Networks," in *Proc. 12th ACM International Conference on Distributed and Event-based Systems*, 2018, pp. 64–75.

[21] F.N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-Level Accuracy With 50x Fewer Parameters And <0.5mb Model Size," *Computer Vision and Pattern Recognition (cs.CV); Artificial Intelligence (cs.AI)*, pp. 1-12, Nov. 2016. DOI: arXiv:1602.07360.

[22] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25 (NIPS 2012).*

[23] A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, P. Jin, S. Zhao, and K. Keutzer, "SqueezeNext: Hardware-Aware Neural Network Design," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 1638-1647. DOI: arXiv:1803.10