IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 11, No. 1, March 2022, pp. 121~128
ISSN: 2252-8938, DOI: 10.11591/ijai.v11.i1.pp121-128

A sound event detection based on hybrid convolution neural network and random forest

Muhamad Amirul Sadikin Md Afendi[1], Marina Yusoff[1,2]
[1]Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Al-Khawarizmi Complex, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

Article Info

Article history:

Received Jun 26, 2021
Revised Dec 23, 2021
Accepted Jan 8, 2022

Keywords:

Convolution neural network
Random forest
Sound event detection
Support vector machine
Wildlife reserve conservation

ABSTRACT

Sound event detection (SED) assists in the detainment of intruders. In recent decades, several SED methods such as support vector machine (SVM), K-Means clustering, principal component analysis, and convolution neural network (CNN) on urban sound have been developed. Advanced work on SED in a rare sound event is challenging because it has limited exploration, especially for surveillance in a forest environment. This research provides an alternative method that uses informative features of sound event data from a natural forest environment and evaluates the CNN capabilities of the detection performances. A hybrid CNN and random forest (RF) are proposed to utilize a distinctive sound pattern. The feature extraction involves mel log energies. The detection processes include refinement parameters and post-processing threshold determination to reduce false alarms rate. The proposed CNN-RF and custom CNN-RF models have been validated with three types of sound events. The results of the suggested approach have been compared with well-regarded sound event algorithms. The experiment results demonstrate that the CNN-RF assesses the superiority with remarkable improvement in performance, up to a 0.82 F1 score with a minimum false alarms rate at 10%. The performance shows a functional advantage over previous methods.

Corresponding Author:

Marina Yusoff
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI)
Al-Khawarizmi Complex, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
Email: marina998@uitm.edu.my

1. INTRODUCTION

Sound event detection (SED) recognizes a sound event's presence using artificial intelligence methods. Applying SED in detecting intrusions for wildlife reserves seems appropriate as the sound of poachers' activities are very distinctive within the natural ambience. Previous works on SED have used various types of machine learning (ML) and deep learning (DL) algorithms such as random forest (RF) and convolutional neural network (CNN). SED solutions were reviewed on an artificial dataset for detecting and classifying acoustic scenes and events [1], [2]. The database comprised isolated sound events with unique sources of background recording, baby crying, gunshots, and glass breaking. The dataset was a mixture of noises and sound events. The researchers used CNN-long short-term memory (CRNN-LSTM) to obtain 93% F1 score [3]. CNN obtained 91% [4], multilayer layer perceptron (MLP-CNN) with 84% [5] and ensemble with 78% [2]. Although these experiments used artificial datasets, they have demonstrated how SED is feasible for industrial applications. The real-world environment sounds are far from noise-free [6]. The SED performances on noisy data can be rather challenging. Sound event's effective detection rate varies n recent

SED experiments and to a certain extent, can be biased to the context [7]. Many sounds overlap on frequency, making it more perplexing to conduct SED. The noises in a forest environment are unique due to the species of animals and plants within the vicinity.

In the SED application, features need to be extracted from the raw audio data to a piece of tailored information. This is done to distinguish sound events more effectively. A common feature used in recent SED studies is mel-log energies (MLE) features that are rich with essential values that contribute to class recognition. The methods explored by the previous research works frequently report that using MLE features with CNN produce good results [8], [9]. These features significantly support models' performance [10]. MLE features are an extraction process that includes fast Fourier transform (FFT). It separates sound frequency within a sound signal. Frequency separation helps in detection as each sound has a specific frequency range. The process would increase the feature determination among the sound events.

CNN is well known for its excellent detection of images [11], [12]. Therefore, CNN is believed to be reliable for SED. CNN uses a large amount of data to work best and overcome challenges such as overfitting, exploding gradient, and class imbalance in the training process [10], [13]. One solution created is by using data augmentation in an urban environment [14], [15]. However, limited research work has been done within the forest environment. The solutions are RF, distance-based [16] and DL methods [11], [12], [17], [18]. The overall performance of these solutions is less than 80% accuracy and has a high false alarms rate. Many solutions have been established mainly in an urban environment with a hybrid approach such as convolution recurrent neural network (CRNN) [19], LSTM-CNN, CNN-support vector machine (SVM) [20]. A study using ensemble methods obtained 85% accuracy on urban rare sound event detection [21]. Recent works on SED with hybrid approaches using CRNN with ensembles achieved 91% accuracy on the artificial and urban datasets [22]. A type of hybrid CNN acts as a feature extractor that improves the feature quality. A hybrid method of CNN and RF also has advantages. CNN-RF's attempts have improved accuracy than earlier methods on pattern recognition tasks on PlantCLEF 2019 dataset [23]. RF algorithm depends on good features like any other algorithms. Without CNN pre-processing high correlated features may affect performance [24]. Hence, the SED field requires more research to be conducted in obtaining the appropriate method of SED solution for surveillance in a forest environment. This research, then should act as an extension to increase the security performance to stop poachers and illegal activities in the forest. The contributions of this study are to provide an alternative solution for SED in the forest environment, to extract the most important features from sound events with a suitable method, to introduce an enhanced solution of hybrid CNN-RF, to provide post-processing thresholding with a minimum false alarm rate of 10%, to evaluate the proposed method using a real dataset and to compare the new solution with other machine learning methods.

This article is organized into several sections. Section 2 presents detailed research method that includes the proposed CNN-RF and post-processing thresholds. Section 3 presents the computational results, comparisons of custom CNN, pre-built CNN, and the relevant discussions. Finally, in section 4, the conclusion and future work are briefly highlighted.

2. PROPOSED MODEL

CNN model is used from a pre-trained model that has been proven to deliver performance for image recognition and the custom CNN produced by the study. The CNN model is based on the VGG16 architecture with the pre-trained weights [11], [12]. Many resources are required to optimized weights. Thus, the use of the trained weights can reduce training time. The customized CNN model used for feature extraction to compare performances of CNN developed from scratch. This is a transfer learning method. Figure 1 shows the flow of the process for the CNN-RF model.

The CNN part of the hybrid model acted as a feature extraction layer. The extracted features would be optimized to be the most useful for the sound classification. The RF performed the prediction from a CNN output. At the end of the process, we applied post-processing to optimize the results. The first step is a sound with a 5-second segment is based applied MLE extraction into MLE features. The CNN model used as a feature extraction method to obtain meaningful features from the MLE features. CNN output was used as the input for the RF model to compute the prediction. The RF ensemble size could be changed to find the best producing parameters. Once the best model was found, the post-processing was performed. The performances between models were compared to analyze the improvements of the hybrid method.

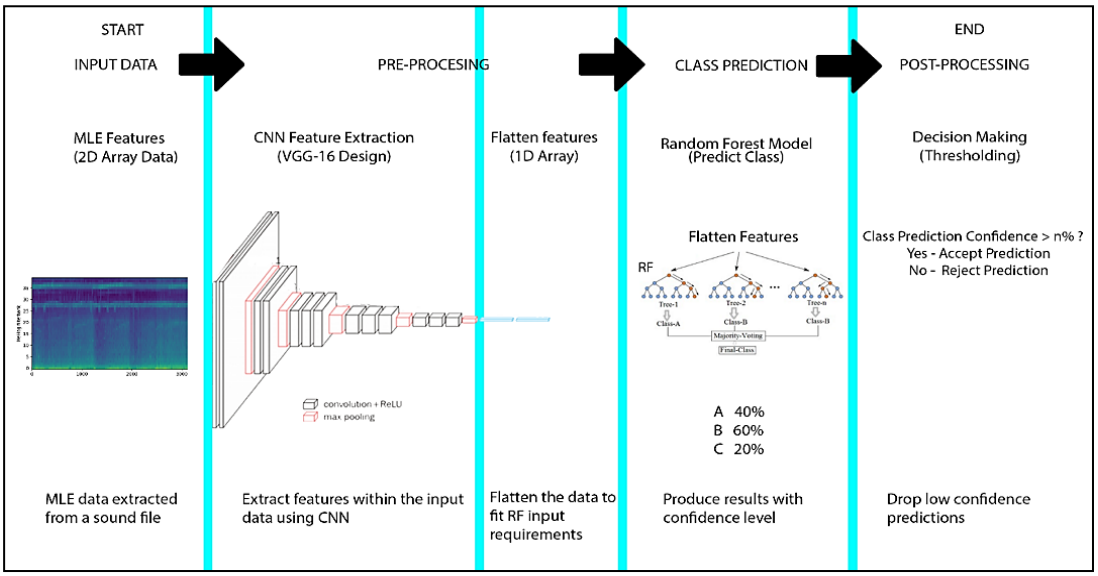

Figure 1. The process flow for the proposed CNN-RF model

3. RESEARCH METHOD

This section provides the explainations on the research steps taken. The steps included data collection, feature extraction, analysis of features, and the classification models of CNN, SVM, RF, and CNN-RF. The steps were done within the respective parameters and configurations.

3.1. Data acquisition

Wildlife reserve intrusion sound detection variables include illegal activities done by poachers. Poachers' equipment detained are mostly axes, machetes, and tools in the forest [25]. These sounds can be the solution to detect incoming threats. These sounds could be good indicators in detecting the presence of poachers in a reserve forest. Therefore, we performed data collection at Taman Negara Endau Rompin, Malaysia. Sound events were emulated to get the sound in the jungle with its current condition and environment. Each location had sound emulations of tree cutting, chainsaw, and vehicle activities. The distance from the sound source was the distance from the source emission point to the recording point in metres. It was scattered into three different distances, approximately at 30 m, 60 m, and 100 m in which were based on an average person's hearing capability [26].

3.2. Sound data feature extraction

MLE was selected as the feature extraction method for its high reliability on rare sound detection in past studies [10], [27], [28]. The steps involved were input signals, Hamming window, FFT, mel-scale filter bank and log. Hamming windows were recommended in this experiment for their properties for the frequency-selective analysis. Hamming windows and the corresponding spectrum form were adapted from [27], [28]. The cepstral features were computed by taking the FFT of the warped logarithmic spectrum. They contained each spectrum band's rate of change [27]. The pth filter bank utilized (1). This showed that f(p) was the middle frequency of the pth filter [29]. This work extended the heatmap representations of the illegal intrusion activities [9], [10].

$$Hm(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{\mathrm{k} - \mathrm{f(m-1)}}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{\mathrm{f(m+1)} - \mathrm{k}}{\mathrm{f(m+1)} - \mathrm{f(m)}}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (1)$$

3.3. Data preparation

For data preparation, the best practices found in recent studies and suitable for CNN training were employed. The data was divided into two subsets: in-sample and out-of-sample data. In-sample data was used

for training and validation of the CNN model. Meanwhile, the 30% out-of-sample data was used for the evaluation of the trained model. The model utilized 70% in-sample data training and 30% validation data. After the model was trained, then it was evaluated using the out-of-sample data. The out-of-sample data was based on both the thick forest and forest roadside environment.

3.4. Evaluation

The algorithms ran on the same machine to avoid any hardware performance inconsistency. The model evaluation metrics used were F1, precision and recall score performed on the out-of-sample data. The prediction results of a ML model were in the form of a collection of TP, TN, FP and FN [30], [31]. The efforts to maintain consistency were to control the hardware and software configurations with Intel Xeon E5 v4 8 Core 16 Threads, Geforce GTX 1080 Graphics Card 8GB GDDRX5, 16 GB DDR4 2666 MHz, and 1TB SATA SSD.

3.5. Thresholding prediction post-processing

Thresholding post-processing on prediction reduced the false alarm by rejecting the low confidence predictions as an additional effort. The determination of predicted class was based on the confidence probability value of each class. By default, a class with the highest value of confidence was selected. The thresholding method was expected to improve performance.

4. COMPUTATIONAL RESULTS AND DISCUSSION

This section presents the computation results of the SED performances. The experiments were conducted on all SED datasets and post-processing measurement on the false alarm rates. A detailed discussion is also provided to elaborate the findings and initiations in this section.

4.1. Model hyperparameters optimization

Several strategies were employed to avoid overfitting. This study used hyperparameters by reducing 250 batch sizes to 125, including a drop out layer of 0.5 on the dense layer, and adjusted the learning rate by decay staircase starting at 1.0 and reduced 90% at each epoch to reduce time to reach early convergence of training data on the early epochs. The design was inspired by the VGG16 model [32], which has half its previous convolutional layer. The new model 32-16-Conv 32 was compared with the previous model and it showed less overfitting. A detailed observation of the results is tabulated in Table 1. The performing model was 32-16-Conv32, 32-32-Conv 32 with early stopping at the 26th epoch with the lowest validation loss while having a considerable loss gap.

Table 1. Performance of custom CNN-RF models

Model	Stop Epoch	Validation Accuracy	Validation Loss	Training Loss	Loss Gap
16-32-Conv 32	26	0.7298	0.7487	0.7155	0.0332
32-32-Conv 32	26	0.8705	0.3824	0.2257	0.1567
32-16-Conv 32	26	0.8526	0.4659	0.5236	0.0577
16-32-Conv 64	21	0.8602	0.367	0.1036	0.2634
32-32-Conv 64	29	0.8975	0.2748	0.1087	0.1661
32-16-Conv 64	28	0.8677	0.3066	0.1759	0.1307

4.2. Computational results of the custom CNN-RF and VGG16-CNN-RF

CNN model of 32-16-Conv 32 implemented in this experiment was the model produced from tuning hyperparameters. The weights trained on this model in layer 32-16 convolutional layers were extracted to be used as a feature extraction layer for the hybrid CNN-RF model. A series of experiments was executed in search of the best parameters for an optimized model with the best performing results. Table 2 demonstrates the performance result of multiple RF models varying in ensemble size, from 10 to 1000. The peak performance of 32-16 CN-RF was achieved at the ensemble size of 500 with an accuracy of 0.7812, F1 of 0.7696, precision 0.7722 and recall of 0.7711.

The results were saturated at 0.7812 at the ensemble size of 300 to 500. The performance started dropping at 1000 ensembles, it seemed that more ensembles were not always better but may cause the model to perform otherwise. The research found that the model of 500 ensembles was the optimized model. The CNN model VGG16 implemented in this experiment was the model acquired from ImageNet [33]. It was found that the ensemble size of 400 trees had produced the best results F1 score of 0.8250, precision 0.8370, and recall 0.8187.

4.3. Post-processing results

After post-processing, the best model maintained a good performance while reducing false alarms. At 75% threshold, the model produced under 10% false alarm rate while maintaining the performance. The threshold reduced all false alarm rates to about 10% while maintaining performance. The post-processing step has improved the model's F1 score by 2.46%, precision by 0.30% and recall reduced by 0.73%. The post-processing improved the overall performance without affecting the performance on F1, precision and recall scores. Figure 2 shows the confusion matrix results of the CNN-RF model. The confusion matrix shows the performance of each class. The results after post-processing reduced the false alarm rate to about 10% and maintained the prediction rate of Hatchet class at 77.5%, Chainsaw 88.9% and Vehicle 88.8%.

Table 2. 32-16-CNN-RF results on different ensemble sizes

Ensemble Size	32-16-CN-RF				VGG16-CNN-RF			
	Accuracy (%)	F1	Precision	Recall	Accuracy (%)	F1	Precision	Recall
10	0.7705	0.7187	0.7220	0.7156	0.7725	0.7781	0.7885	0.7724
20	0.7761	0.7454	0.7548	0.7460	0.7855	0.7982	0.8072	0.7931
50	0.7759	0.7596	0.7610	0.7627	0.7933	0.8158	0.8251	0.8109
100	0.7755	0.7614	0.7626	0.7643	0.7964	0.8194	0.8299	0.8138
200	0.7793	0.7700	0.7726	0.7717	0.7962	0.8200	0.8312	0.8143
300	0.7784	0.7693	0.7716	0.7708	0.7965	0.8218	0.8331	0.8159
400	0.7802	0.7692	0.7715	0.7708	0.7969	0.8250	0.8370	0.8187
500	0.7812	0.7696	0.7722	0.7711	0.7968	0.8240	0.8358	0.8180
1000	0.7791	0.7698	0.7723	0.7710	0.7976	0.8229	0.8332	0.8177

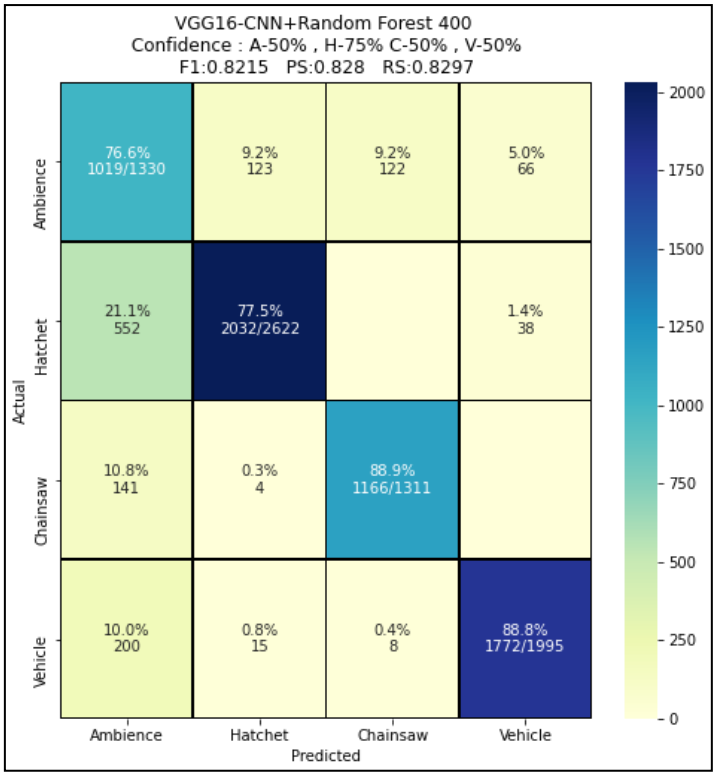

Figure 2. CNN-RF after post-processing confusion matrix

The results show a significant improvement for RF and CNN-RF, as demonstrated in Table 3. The CNN-RF had 12.55% F1 score. The ensemble size was then lowered to produce the best results. It was demonstrated that it was more efficient as it used a smaller ensemble size. The CNN-RF achieved was less than 10% FP rate at 75% Hatchet threshold than 85% on RF.

The accepted threshold value setting was the value when FP rates reached approximately 10%. The FP rate was essential for the evaluation of the algorithm. Another aspect to consider was that the ensemble size contributed to more computational time requirement. Small ensembles size should be relatively more efficient. However, this study did not include an in-depth evaluation on this aspect.

The performance of each model was compared before the post-processing and after thresholding. Table 4 presents the comparison in performance between models after the post-processing prediction. The study listed the scores in which the thresholding effort produced lower than 10% for the false alarm rate. Based on the obtained FP rate, the misclassification of classes could be identified.

Based on the observation, the average VGG16-CNN-RF model performed good results with an F1-score 0.8215, but it also reduced the performance of individual class accuracy, the Chainsaw and Hatchet class detections to 88.9% and 88.8% respectively. The 32-16-CNN-RF model acquired the best results in detecting Chainsaw and Vehicle class with 98.8% and 92.2% accuracy. The combination of CNN and RF could improve the individual class accuracy of the hatchet. The VGG16 was well-trained, but within the domain of images, perhaps a well-trained CNN for SED could be established using a more variety of real datasets in the future.

Table 3. Comparison between VGG16-CNN-RF and RF

Algorithm	Key performance scores			Threshold (%)	Individual class performance		
	F1	Precision	Recall		Hatchet	Chainsaw	Vehicle
RF	0.696	0.761	0.754	0.85	27.2	99.3	85
VGG16-CNN-RF	0.822	0.828	0.83	0.75	77.5	88.9	88.8
Difference	+0.126	+0.067	+0.076	-0.10	+50.3	-11.6	+3.0

Table 4. Comparison between models with post-processing

Algorithm	Key performance scores			False alarm rate (approx. 10%)	Individual class performance accuracy (%)		
	F1	Precision	Recall	Reliable Threshold (%)	Hatchet	Chainsaw	Vehicle
RF	0.6960	0.7606	0.7535	85	27.z2	99.3	85.0
32-16-Conv CNN-RF	0.6304	0.7112	0.6965	88	12.2	98.8	92.2
VGG16-CNN-RF	0.8215	0.8280	0.8297	75	77.5	88.9	88.8
32-16-Conv 32 CNN	0.7762	0.8007	0.8018	80	80.0	99.8	86.4
SVM	0.5694	0.6515	0.6296	92	20.7	84.4	71.3

4.4. Discussion

The research explored three methods, namely CNN, RF and SVM. A proposed method of CNN-RF has been exclusively established and studied for SED solutions in a forest environment. The results of the CNN-RF model showed considerable improvement with the F1 score of 0.7762. However, the loss function of multi-class cross-entropy was 0.42, which demonstrated that it needed more data to improve prediction quality. In contrast, the RF results were further enhanced with the CNN-RF model. The CNN-RF has shown some improvements as the CNN was used as the feature extractor. Besides, the RF was employed as the hypothesis of this combination that may improve performance. This study believes that an image similar to the one of hotspots has emerged based on the MLE pattern analysis done on the collected sound. The CNN 2D layer extracts more spatial features allowing the RF to improve performance with the tuned feature by the hybrid portion of CNN. The observation conducted has found an anomaly of difference in sound event class performance, especially at the Hatchet class. It always shows a lower detection regardless of any models. The nature hatchet sound is different compared to others. Instead of a long sustaining event like the others, it is in multiple bursts. The results show that the model can detect the hatchet event at 77% accuracy and a high FP rate at 30%. The other sound events provide better results of more than 85% and low FP under 10%. The thresholding method can be optimized on individual classes tailored to their respective difficulty in reducing FP.

Each model suffers a high degree of false prediction rate, confuses between ambience event and intruder event. The false prediction is not aligned with the intended purpose of the research for a security surveillance system. Hence, a post-processing layer is considered mandatory. The earliest result is considered too loose with false prediction rates. Thresholding post-processing is applied to the point of approximately 10% false prediction on any intruder events. It is not a good prediction with low confidence of about 51% over the other 49% [34]. Hence, increasing the threshold will avoid this problem. The variable threshold level is optimized until the target of 10% false detection rate is achieved. CNN and CNN-RF can be considered reliable for security in wildlife reserves. However, the findings do not apply in all of the SED cases. Further research is required to set the foundation to implement the use of SED in a vast area of surveillance in forests.

5. CONCLUSION

In conclusion, this study has discovered that many algorithms and techniques in solving SED have feasible application in industries. The CNN-RF has been proven to demonstrate an overall improvement in performance. It has also been discovered that it requires less configuration and optimization efforts due to the capabilities of CNN transfer learning. RF is recommended to be a suitable classifier in the SED task for forest environment compared to others because it is a hybrid approach in tackling SED in the domain. A post-processing method of thresholding has been applied to the prediction results in reducing the FP rate. Thresholding is reduced FP rate from 30% to 10% with an accuracy penalty between 94.6% and 77.5% on Hatchet class. The VGG16 based CNN-RF model and thresholding combination have enhanced performance for surveillance applications with 80% accuracy average and less than 10% of FP rate. Recommendations for subsequent research are to investigate the use of noise cancellation to isolate the featured sound events, to make use of the advantages of other deep learning methods such as Generative Adversarial Network and Attention CNN and to acquire a more optimal thresholding degree on the post-processing threshold. In addition, more experiments with real-world sound samples to understand the contributing factors such as different sound events, environment, noise, location, and weather can also be conducted.

ACKNOWLEDGEMENTS

The authors would like to thank the Research Management Center, Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia for providing essential support and knowledge for this research.

REFERENCES

[1] A. Dang, T. H. Vu, and J.-C. Wang, "Deep learning for DCASE 2017 challenge," *Detection and Classification of Acoustic Scenes and Even, t*2017.
[2] A. Mesaros *et al.*, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.
[3] A. S. Md Afendi, M. Yusoff, and M. Omar, "Mel-log energies analysis of authentic audible intrusion activities in a Malaysian forest," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 2, Apr. 2020, doi: 10.11591/eei.v9i2.2091.
[4] A. S. M. Affendi and M. Yusoff, "Review of anomalous sound event detection approaches," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 3, pp. 264–269, Dec. 2019, doi: 10.11591/ijai.v8.i3.pp264-269.
[5] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Genetics*, vol. 19, no. S1, Art. no. 65, Sep. 2018, doi: 10.1186/s12863-018-0633-8.
[6] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, and D. Evans, "Revisiting membership inference under realistic assumptions," *Computer Science*, May 2020, Available: http://arxiv.org/abs/2005.10881.
[7] C. Tian, Y. Xu, and W. Zuo, "Image denoising using deep CNN with batch renormalization," *Neural Networks*, vol. 121, pp. 461–473, Jan. 2020, doi: 10.1016/j.neunet.2019.08.022.
[8] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604–624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.
[9] E. Cakır and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," *Deep Neural Networks for Sound Event Detection*, vol. 12, 2019.
[10] F. Demir, M. Turkoglu, M. Aslan, and A. Sengur, "A new pyramidal concatenated CNN approach for environmental sound classification," *Applied Acoustics*, vol. 170, Art. no. 107520, Dec. 2020, doi: 10.1016/j.apacoust.2020.107520.
[11] F. J. Harris, "Multirate FIR filters for interpolating and desampling," in *Handbook of Digital Signal Processing*, Elsevier, 1987, pp. 173–287.
[12] H. Lim, J. Park, K. Lee, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," *Detection and Classification of Acoustic Scenes and Events*, 2017.
[13] H. Phan, M. Krawczyk-Becker, T. Gerkmann, and A. Mertins, "DNN and CNN with weighted and multi-task loss functions for audio event detection," Aug. 2017. Available: http://arxiv.org/abs/1708.03211.
[14] J. G. Selman and N. Demir, "Automatic detection for acoustic monitoring of wild animals," 2019. Available: http://ilpubs.stanford.edu:8090/1166/.
[15] J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparison of deep learning methods for environmental sound detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 126–130, doi: 10.1109/ICASSP.2017.7952131.
[16] K. Chung, "Perceived sound quality of different signal processing algorithms by cochlear implant listeners in real-world acoustic environments," *Journal of Communication Disorders*, vol. 83, Art. no. 105973, Jan. 2020, doi: 10.1016/j.jcomdis.2019.105973.
[17] K. K. Lella and A. Pja, "Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice," *AIMS Public Health*, vol. 8, no. 2, pp. 240–264, 2021, doi: 10.3934/publichealth.2021019.
[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014. Available: http://arxiv.org/abs/1409.1556.
[19] K. Wang, L. Yang, and B. Yang, "Audio event detection and classification using extended R-FCN approach," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 128–132, 2017.
[20] M. Z. Alom *et al.*, "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, vol. 8, no. 3, Art. no. 292, Mar. 2019, doi: 10.3390/electronics8030292.
[21] P. Patel and A. Thakkar, "The upsurge of deep learning for computer vision applications," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 538–548, Feb. 2020, doi: 10.11591/ijece.v10i1.pp538-548.

[22] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," Jan. 2021. Available: http://arxiv.org/abs/2101.02919.
[23] R. P. Reynolds, W. L. Kinard, J. J. Degraff, N. Leverage, and J. N. Norton, "Noise in a laboratory animal facility from the human and mouse perspectives.," *Journal of the American Association for Laboratory Animal Science: JAALAS*, vol. 49, no. 5, pp. 592–597, Sep. 2010. Available: http://www.ncbi.nlm.nih.gov/pubmed/20858361.
[24] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 86–90, doi: 10.1109/ICASSP40776.2020.9054478.
[25] S. Banerjee and R. Pamula, "Random forest boosted CNN: an empirical technique for plant classification," in *Advances in Intelligent Systems and Computing*, Springer Singapore, pp. 251–261, 2020.
[26] S.-H. Jung and Y.-J. Chung, "Performance analysis of the convolutional recurrent neural network on acoustic event detection," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1387–1393, Aug. 2020, doi: 10.11591/eei.v9i4.2230.
[27] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 131–135, doi: 10.1109/ICASSP.2017.7952132.
[28] S. Joshi, D. K. Verma, G. Saxena, and A. Paraye, "Issues in training a convolutional neural network model for image classification," in *Communications in Computer and Information Science*, Springer Singapore, pp. 282–293, 2019.
[29] S.-Y. Jung, C.-H. Liao, Y.-S. Wu, S.-M. Yuan, and C.-T. Sun, "Efficiently classifying lung sounds through depthwise separable CNN models with fused STFT and MFCC features," *Diagnostics*, vol. 11, no. 4, Art. no. 732, Apr. 2021, doi: 10.3390/diagnostics11040732.
[30] P. Tzirakis, A. Shiarella, R. Ewers, and B. W. Schuller, "Computer audition for continuous rainforest occupancy monitoring: the case of bornean gibbons' call detection," in *Interspeech 2020*, Oct. 2020, pp. 1211–1215, doi: 10.21437/Interspeech.2020-2655.
[31] S. L. Ullo, S. K. Khare, V. Bajaj, and G. R. Sinha, "Hybrid computerized method for environmental sound classification," *IEEE Access*, vol. 8, pp. 124055–124065, 2020, doi: 10.1109/ACCESS.2020.3006082.
[32] W. Liu and J. A. Zagzebski, "Trade-offs in data acquisition and processing parameters for backscatter and scatterer size estimations," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 57, no. 2, pp. 340–352, Feb. 2010, doi: 10.1109/TUFFC.2010.1414.
[33] Z. Liu and S. Li, "A sound monitoring system for prevention of underground pipeline damage caused by construction," *Automation in Construction*, vol. 113, Art. no. 103125, May 2020, doi: 10.1016/j.autcon.2020.103125.
[34] Z. S. Bojkovic, B. M. Bakmaz, and M. R. Bakmaz, "Hamming window to the digital world," *Proceedings of the IEEE*, vol. 105, no. 6, pp. 1185–1190, Jun. 2017, doi: 10.1109/JPROC.2017.2697118.

BIOGRAPHIES OF AUTHORS

Muhamad Amirul Sadikin Md Afendi received Bachelor of Information Technology (Hons.) Intelligent System Engineering 5th July 2019. Currently, he is a Master student at Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia. He is a freelance fullstack web-system developer since 2018, working on various projects creating workflow management software. Projects previously ventures into microcontroller programming, applying ML with microcomputers, electronics hardware prototyping, and webserver cloud computing. He can be contacted at email: ai.amirul.sadikin@gmail.com.

Marina Yusoff is currently a senior fellow researcher at the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) and Associate Professor of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Malaysia. She has a Ph.D. in Information Technology and Quantitative Sciences (Intelligent Systems). She previously worked as a Senior Executive of Information Technology in SIRIM Berhad, Malaysia. She is the most interested in multidisciplinary research, artificial intelligence, nature-inspired computing optimization, and data analytics. She applied and modified AI methods in many research and projects, including deep learning, neural network, particle swarm optimization, genetic algorithm, ant colony, and cuckoo search for many real-world problems and industrial projects. Her recent projects are data analytic optimizer, audio and image pattern recognition. She has many impact journal publications and contributed as an examiner and reviewer to many conferences, journals, and universities' academic activities. She can be contacted at email: marina998@uitm.edu.my.