Applied Acoustics 175 (2021) 107829

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers

Mahmoud A. Alamir

College of Science and Engineering, Flinders University, Clovelly Park, Adelaide, SA 5042, Australia

ARTICLE INFO

Article history:
Received 20 August 2020
Received in revised form 1 October 2020
Accepted 25 November 2020
Available online 16 December 2020

Keywords:
Artificial intelligence
Noise management
Environmental noise classification
Acoustic scene classification
Late fusion

ABSTRACT

Recent evidence suggests that convolutional neural networks (CNNs) can model acoustic scene classification (ASC) with high accuracy. Ensemble classifiers have also shown high accuracy in different machine learning areas. However, little is known about fusion models between CNNs and different ensemble classifiers for ASC. This study presents an enhanced CNN classification model using the late fusion between CNNs and ensemble classifiers to predict different classes of acoustic scenes. A CNN model was first built to classify fifteen acoustic scene environments. Different ensemble classifier models were then used for this classification problem. Late fusion of CNN and ensemble classifier models was then applied. The results showed that late fusion models have higher classification accuracy, as compared to individual CNN or ensemble classifier models. The best model was obtained by fusion of the CNN and discriminant random subspace classifier with an increase in the average accuracy of 10% as compared to the average accuracy of the CNN model. When compared with previous research on ASC, the late fusion model between CNN and ensemble classifiers showed higher accuracy. Therefore, this method has robust applicability for future ASC problems.

1. Introduction

1.1. Background

Acoustic scene classification (ASC) is the way to classify different environments depending on their sound characteristics. The scene, in this context, refers to the acoustic environment summarised in one situation such as "restaurant" or "office". Acoustic scenes could be pre-recorded or live streaming audio [1]. ASC plays an important role in many areas, such as context awareness in smart devices, hearing aids, robots, and many other applications [2]. However, there is a need for high-performance ASC models [3]. Therefore, many algorithms and methods have been developed to achieve accurate ASC models.

1.2. Previous methods for ASC

There have been many methods for ASC, mostly using CNNs [2,4]. However, early fusion CNN models showed high accuracy for ASC. Fusion means combining one or more characters at the same time. In terms of modelling, fusion can be classified into early and late fusion. Early fusion could refer to combining the features using more than one method or other concepts such as refining frequency resolution before beginning model training. Recently, early fusion models have extensively been used for ASC. For example, Yang et al. [5] used multistage feature extraction fusion for ASC. Su et al. [6] also used aggregated feature extraction for ASC. Zhang et al. [7] used fine-resolution frequency for feature selection of ASC. Mulimani et al. [8] also used fisher vector for feature extraction of ASC.

1.3. Previous results of late fusion models

Late fusion refers to combining the results of different models after building each model separately [9]. Recently, late fusion models were used in many areas because of their higher predictability as compared to individual models. For example, they have shown higher predictability than early fusion models when used for semantic video analysis [10]. Late fusion can be achieved by combining CNN model with other models such as SVM or different CNN models with different feature extraction methods [11]. Recently, it was also used for emotion recognition for audio-visual data [12,13]. They were also used for recognising human activity [14]. However, the use of late fusion models for ASC has not been applied before between CNN and different ensemble classifier models for ASC problems.

E-mail address: mahmoud.alamir@flinders.edu.au

https://doi.org/10.1016/j.apacoust.2020.107829
0003-682X/

1.4. Study aims

Most studies optimising ASC models are based on the early fusion of feature characteristics before using them in CNN models. It is hypothesised that late fusion of different models could yield higher predictive power, as compared to when using only one model. Therefore, this study proposes a late fusion model between CNN and ensemble classifier models. Different ensemble classifiers are studied and their accuracy, when fused with CNN, is also presented. The results help to improve ASC predicted accuracies.

2. Acoustic scene data

2.1. Data source

The dataset of TUT Acoustic scenes 2017 challenge was used [15]. A description of acoustic scenes included in the dataset can be found through http://www.cs.tut.fi/sgn/arg/dcase2016/acoustic-scenes#library. The dataset consists of various acoustic scenes recorded from distinct locations. Each acoustic scene has 312 segments for training noise samples and 108 for testing noise samples. For each original recording location, a 3–5-minute-long audio recording was captured.

2.2. Acoustic scene types

Acoustic scenes were classified as follows: Bus - travelling by bus in the city (vehicle), Cafe / Restaurant - small cafe/restaurant (indoor), Car - driving or travelling as a passenger, in the city (vehicle), City centre (outdoor), Forest path (outdoor), Grocery store - medium size grocery store (indoor), Home (indoor), Lakeside beach (outdoor), Library (indoor), Metro station (indoor), Office - multiple persons, typical workday (indoor), Residential area (outdoor), Train (travelling, vehicle), Tram (travelling, vehicle), Urban park (outdoor).

2.3. Recording acoustic scenes

All included acoustic scenes were recorded in different locations (i.e. different parks, streets, homes). The sound was recorded using "Soundman OKM II Klassik/studio A3", electret binaural microphone and a "Roland Edirol R-09" wave recorder with a sample rate of 44.1 kHz and a resolution of 24-bit. The microphones were designed to be similar to headphones when used in-ears [16,17]. This allowed recorded sounds to be similar to the sound in the auditory system [18,19].

2.4. Post-processing audio file dataset

The recorded data was post-processed to make for the privacy of people and any possible errors in recordings. For data recorded in private places, consent forms were filled from all people occupying these places. Consent forms were not required for data recorded in public places [20–22]. However, the content with privacy segments was removed. Any data with failure either because of microphones and/or signal distortions were also removed. After cleaning data, remaining data files were segmented into 10-second data files. Table 1 summarises the included 10-second acoustic scene segments.

3. The proposed method

3.1. Overview

Fig. 1 shows the proposed late fusion model procedures. Data was first entered and was then split into 10-sec segments. Feature extraction was then done by applying Mel-spectrograms to convolutional neural networks (CNNs) and wavelet scattering for ensemble classifiers. Hyper-parameter tuning was then done for CNN and ensemble classifier models separately. The fusion of CNN and ensemble classifier models was then applied to maximise the accuracy obtained. Each model (i.e. the CNN, ensemble classifier and fusion models) was then used for class prediction and its accuracy was evaluated. Different ensemble classifiers were used. Therefore, these procedures were repeated with each ensemble classifier. All these procedures and functions to execute them were done through MATLAB 2020a.

3.2. Convolutional neural networks

3.2.1. Data augmentation

The DCASE 2017 dataset contained a relatively small number of acoustic recordings, and the development set and evaluation set were recorded at different specific locations. As a result, it is easy to overfit to the data during training. To fix this problem, data was augmented in two ways.

The first augmentation method was done by splitting each 10-second sample into 10 one-second samples. The split procedure made CNN easier to train and avoid overfitting to any acoustic event. It also ensured the relative combinations of the training data events. Data was also augmented to increase the training stage accuracy.

The second augmentation method was by implementing mix-ups. In a mix-up, the dataset is augmented by mixing features of two different classes. When features are mixed, labels are mixed in equal proportion [23]. When training proposed CNN model, labels were drawn from probability distribution instead of mixed labels. Each spectrogram was mixed with a spectrogram of a different label with lambda set to 0.5. Original and mixed datasets are combined for training. These procedures can be described in Eqs. (1) and (2), where $\bar{x}$ is spectrogram features extracted from a random available spectrogram (x_j) added to the augmented sample (x_i) and $\bar{y}$is the label of randomly set by lambda.

$$\bar{x} = \lambda x_i + (1-\lambda)x_j \tag{1}$$

$$\bar{y} = \lambda y_i + (1-\lambda)y_j \tag{2}$$

3.2.2. The architecture of the proposed convolutional neural network

The convolutional neural network (CNN) architecture was based on the design from [2]. However, the number of learning

Table 1
A summary of the number of training and testing noise samples included in this study.

Type of acoustic scene	Beach	Bus	Cafe/ restaurant	Car	City centre	Forest path	Grocery store	Home	Library	Metro station	office	park	Residential area	Train	Tram
No. of noise samples [Training]	312	312	312	312	312	312	312	312	312	312	312	312	312	312	312
No. of noise samples [Testing]	108	108	108	108	108	108	108	108	108	108	108	108	108	108	108

cycles and training processes were modified as explained in the following sections.

3.2.3. Feature extraction

Mel Spectrograms were used to transform the audio files into frequency-domain representation. The characteristics of these spectrograms were as follows: window length = 2048; samples per-hop = 1024; samples overlap = 1024; FFT Length = 4096; number of bands = 128. Fig. 2 shows a typical example of the extracted spectrograms for each 10-sec sample after augmentation by splitting it into 10 one-seconds audio files.

3.2.4. Convolutional neural network training

The Bayesian optimisation [24] was used to obtain hyperparameters, which were as follows: minimum batch size = 128; momentum = 0.9, L2 Regularization = 0.005, maximum epochs = 15, learn rate schedule = piecewise.

To speed up processing, Mel spectrograms for all noise files in the datastores were extracted using tall arrays. Tall arrays remain uncalculated until calculations are requested using the gather function. These delayed calculations help to work with large data quickly. When requested, MATLAB combines these queued calculations to take the minimum passes through the data.

3.2.5. Convolutional neural network accuracy evaluation

A probability-weighted averaged on the one-second segments was used to predict the scene for each 10-second noise sample in the testing dataset. For each 10-second noise sample, the maximum relative weight prediction was used and labelled to the corresponding predicted acoustic scene.

3.3. Ensemble classifiers

Wavelet scattering has been shown in [25] to provide a good representation of acoustic scenes. Therefore, it was used for extracting the features for the ensemble classifier training. The invariance scale and quality factors were determined through trial and error.

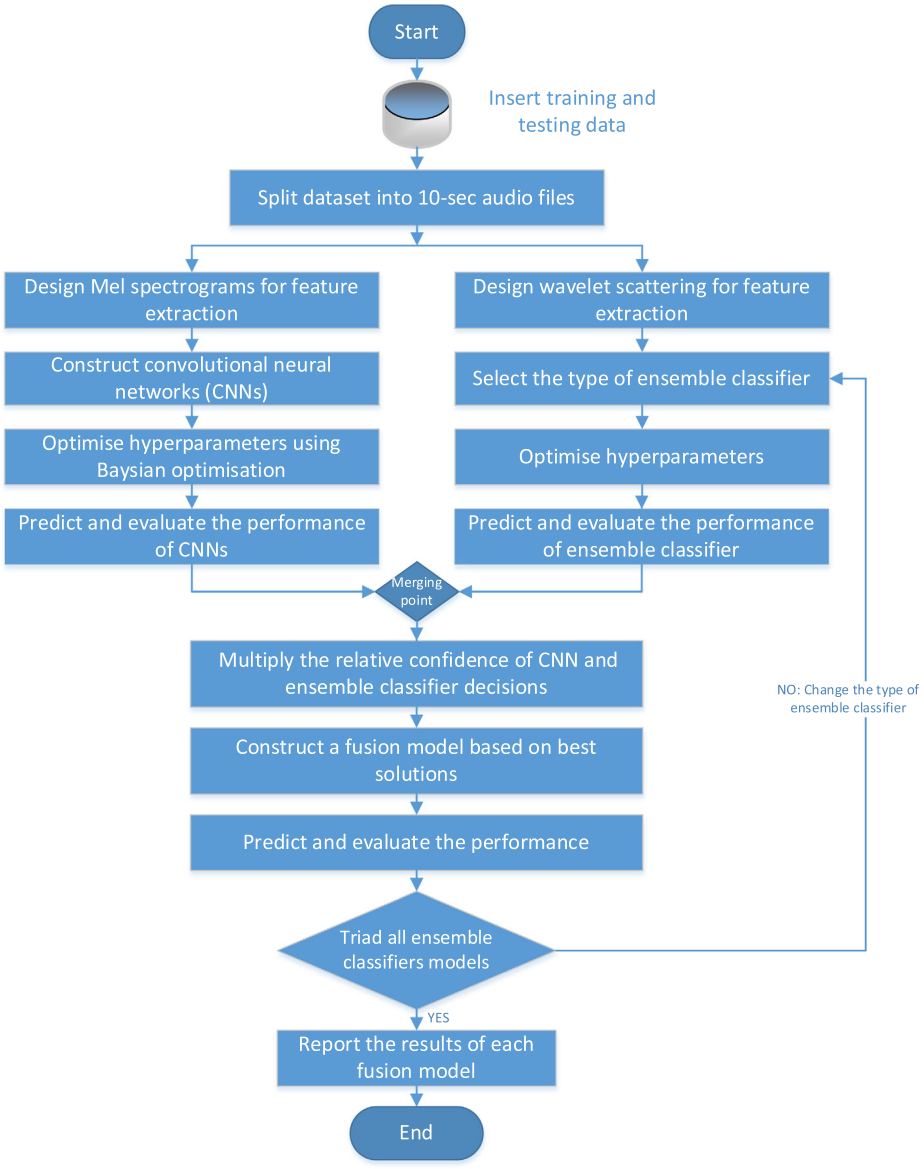


Fig. 1. The proposed late fusion model between CNN and ensemble classifier models. The dataset for each model is augmented firstly and each model is then optimised before the late fusion occurs in the last step.

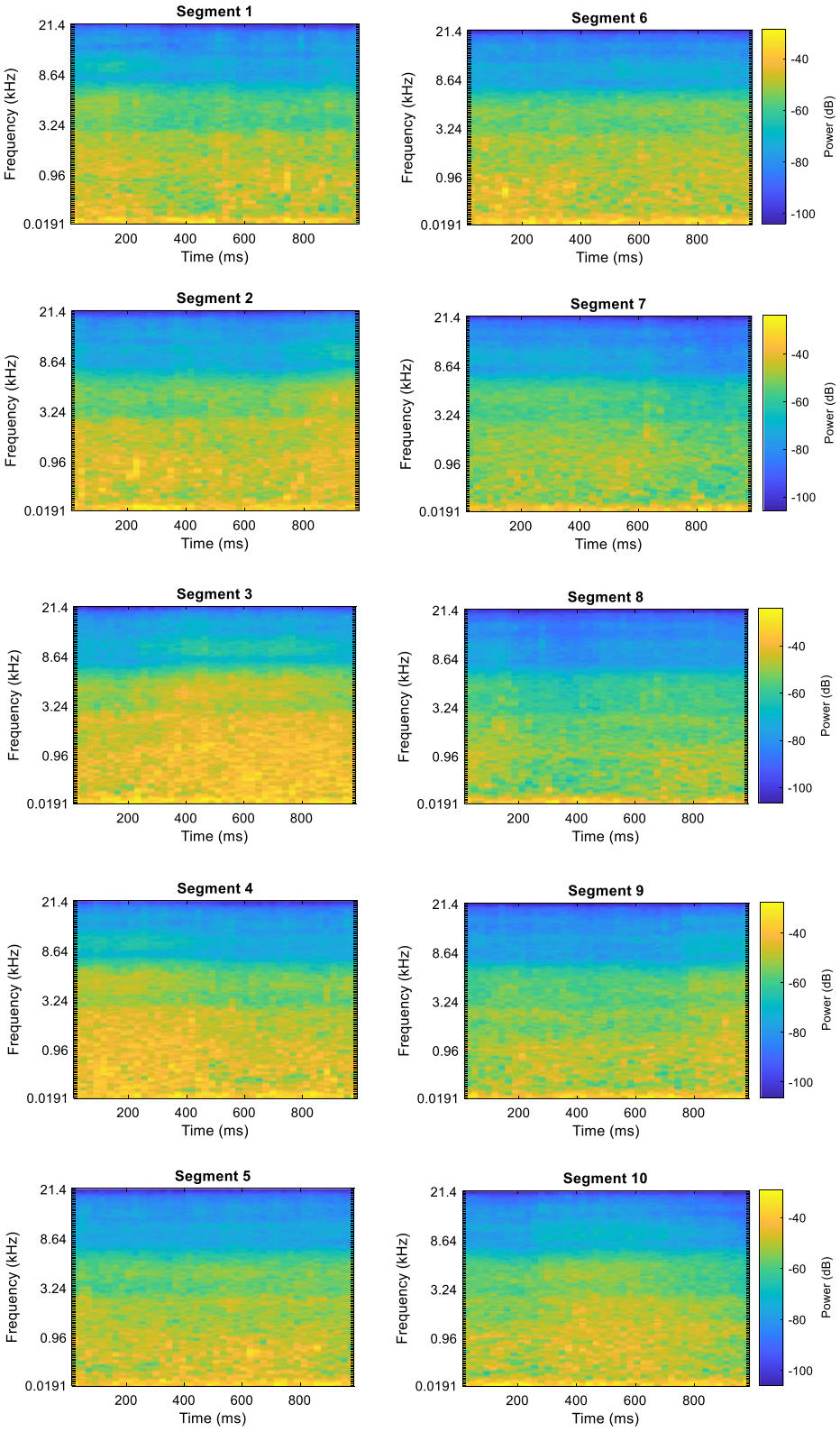


Fig. 2. Mel spectrograms representing the input to the convolutional neural networks (CNNs) of a 10-sec audio file augmented by splitting it into 10 one-second audio files.

Fig. 4. The training progress showing the accuracy and loss of convolutional neural network (CNN) model for each epoch and iteration. Iterations are shown in the horizontal axis, while the y-axis represents accuracy in the upper figure (blue and light blue lines) and loss in the lower figure (red and light red lines). Light colour lines represent smoothed processes, while other lines represent original training procedures. Maximum iterations were set to 5205 (347 iterations per epoch). The shaded vertical rectangles represent the 15 epochs required for the training process. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

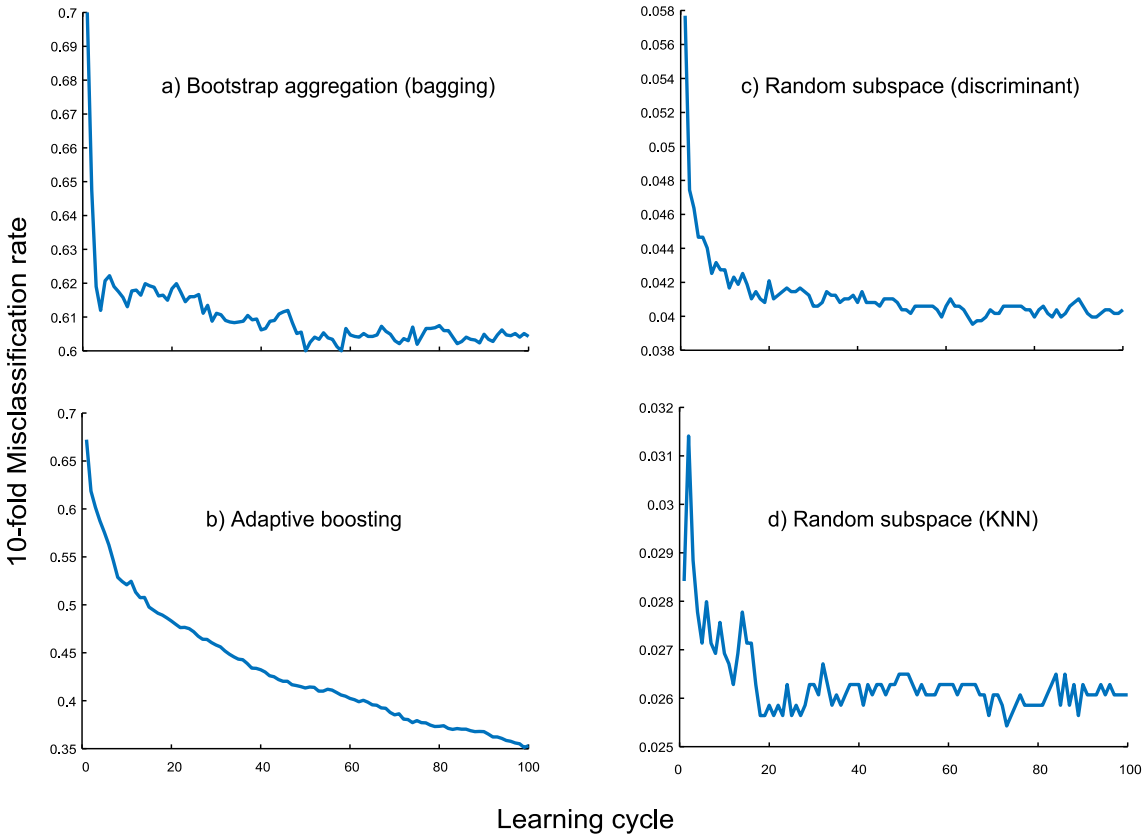
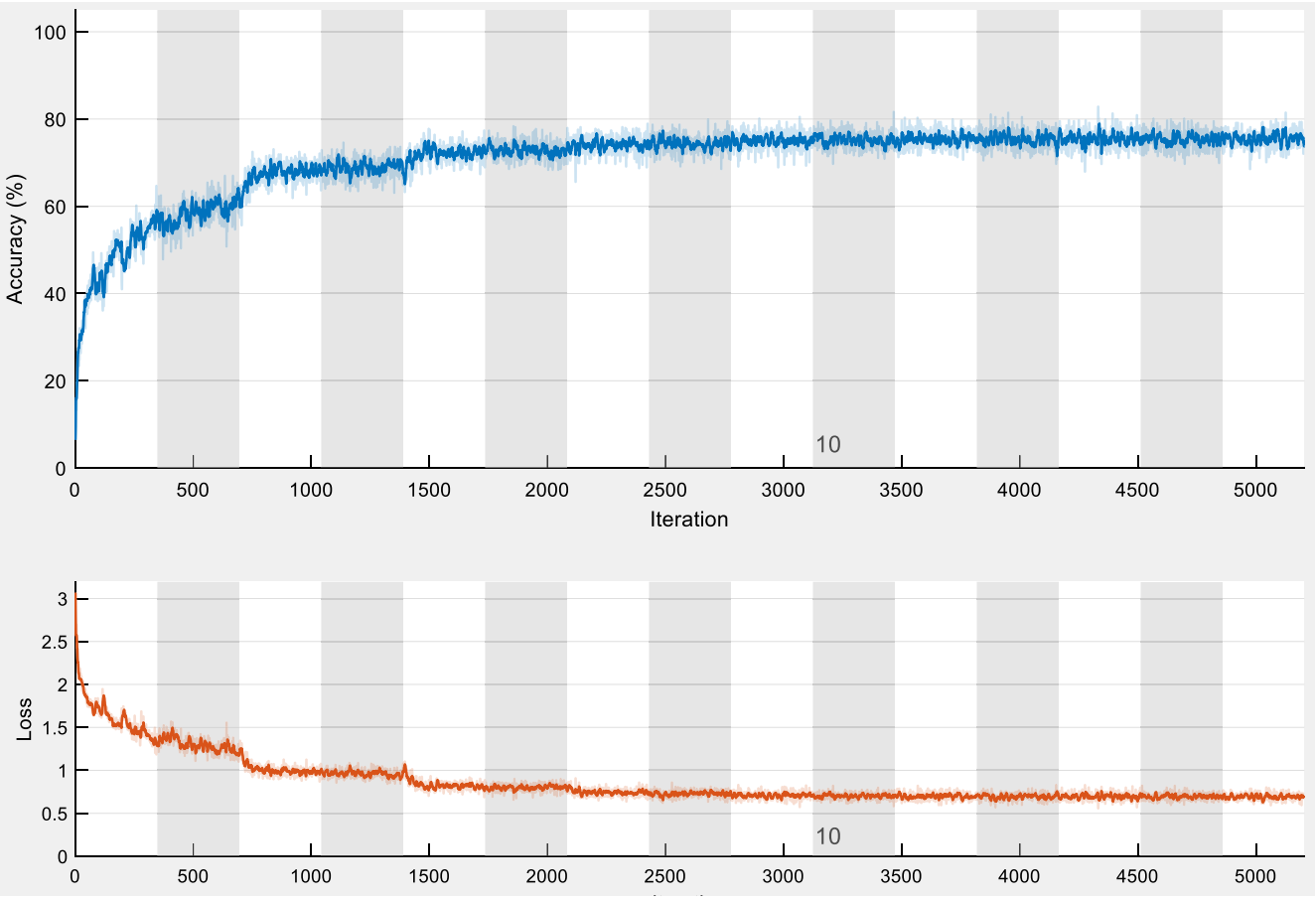
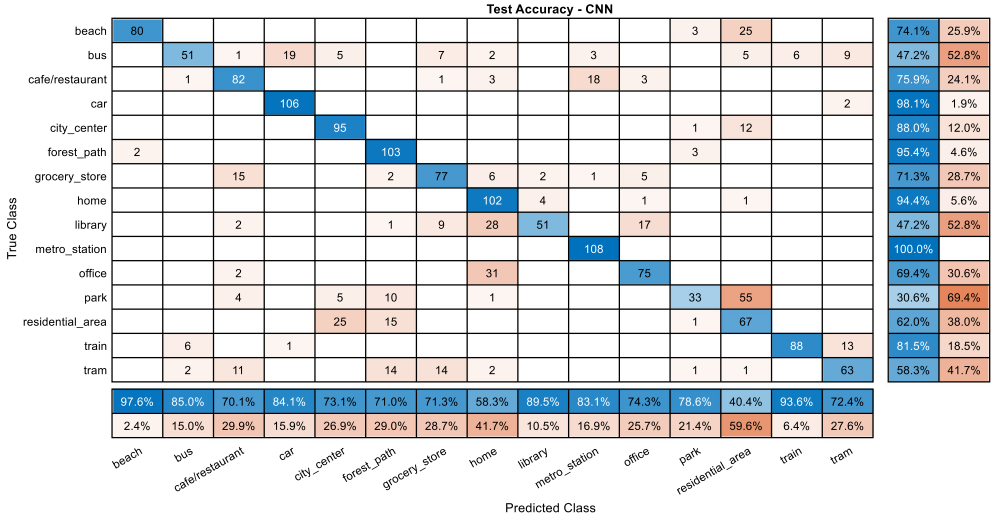
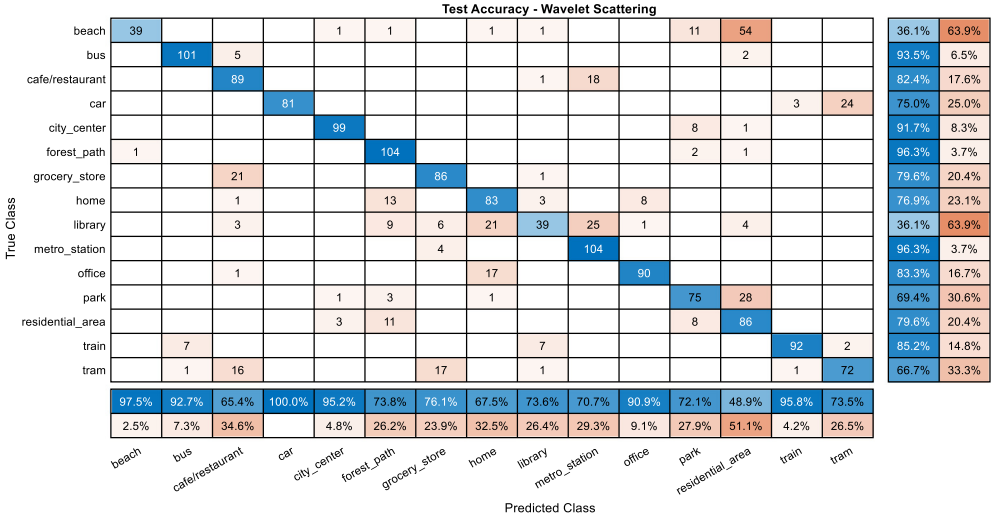


Fig. 3. Examples of hyperparameter optimisation (learning cycle) of some ensemble classifiers. The number of learning cycles that gives the lowest 10-fold misclassification rate was chosen for each classifier.

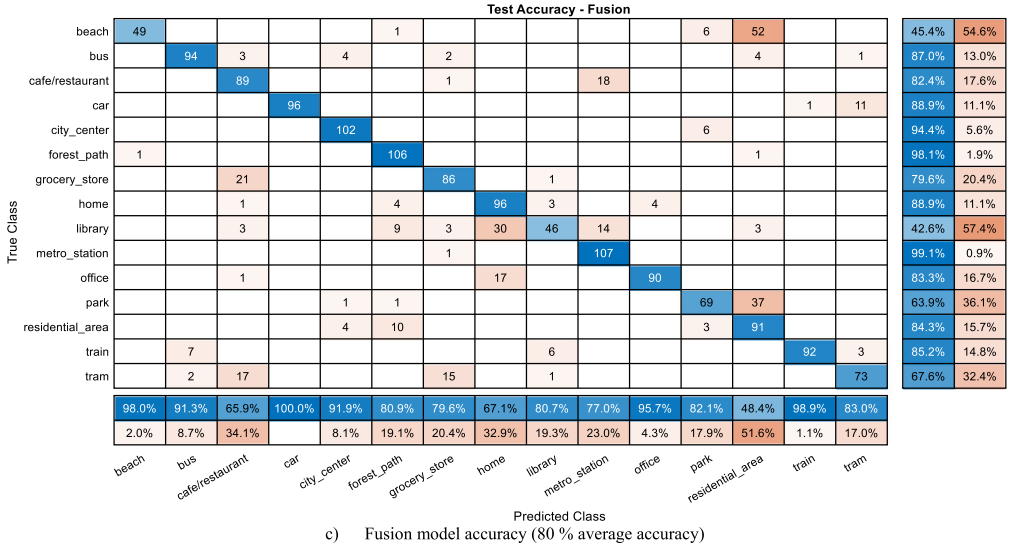




a) CNN model accuracy (72.9 % average accuracy)



b) Random subspace classifier accuracy (76.5 % average accuracy)



c) Fusion model accuracy (80 % average accuracy)

Fig. 5. The confusion matrix of different acoustic scenes for a) CNN model, b) Random subspace discriminant classifier models and c) Fusion model between CNN and random subspace classifier. The number in each cell represents how many times the class shown in the vertical axis was predicted as the class in the horizontal axis. Correct predictions are labelled by blue, while false predictions can be seen by red. Dark colours represent that there were many predictions in these classes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Average accuracy of different ensemble classifiers and their fusion with the CNNs.

	Bootstrap aggregation (Bagging)	Adaptive boosting	Random subspace classifier [discriminant]	Random subspace classifier [KNN]	Linear programming boosting	Random under-sampling boosting	Totally corrective boosting
Model predictive accuracy (%)	67.2	41.3	76.5	50.1	48.3	49.7	40.4
Fusion with CNN predictive accuracy	74.2	69.8	80	57.6	73.1	73.4	62.1

Table 3
The results of the different best CNN methods on DCASE 2017, as compared to the current study. Average accuracy was commonly reported in these studies, so it has been used for comparison.

Study	Weiping et al. [28]	Hyder et al. [29]	Lehner et al. [30]	Park et al. [31]	Piczak [32]
Method	Early feature fusion of spectrogram and Constant-Q-Transform (CQT) [CNN-SQT]	Spectrogram Image Features (SIF) [CNN-SIF]	I-vectors and CNNs	CNN using double image features	CNN with a frequency resolution
Average accuracy	74.8	74.1	73.8	72.6	70.6
Current proposed model accuracy improvement	7%	8%	8%	10%	13%

Seven different ensemble classifiers were used, and their performance was compared. These classifiers included aggregation (bagging), adaptive boosting, random subspace classifier using KNN and discriminant for feature extraction, linear programming boosting, random under-sampling boosting and totally corrective boosting. MATLAB- defined functions were used to build the ensemble classifiers.

The maximum number of decision splits per tree and number of learning cycles was tuned for each ensemble classifier. For example, an optimisation process was done for ensemble hyperparameter (learning cycles) as shown in Fig. 3. This was based on cumulative 10-fold cross-validated misclassification rates. The cumulative loss allowed monitoring loss with accumulating learners in ensembles. The learning cycle number with minimum misclassification rate was used for each ensemble. The number of minimum 10-fold misclassification errors was different for different classifiers as shown in Fig. 3.

The scattering coefficients were obtained for the scattering decomposition framework and then averaged over 10-second noise samples. For each 10-second noise sample, calling "predict" in MATLAB returns the labels of the corresponding predicted location and the relative confidence in the decision [26,27].

3.4. Fusion model

The fusion model makes use of convolutional neural networks (CNNs) using Mel-spectrograms and ensemble classifiers using wavelet scattering. Roughly an equal overall accuracy can be obtained from CNN and ensemble classifiers; however, each model can outperform the other for predicting particular acoustic scenes. Therefore, merging both models could make use of the advantages of each model individually.

To increase overall accuracy, CNN and ensemble classifier results using late fusion were merged. For each 10-second noise samples, using "predict" function for wavelet classifier and CNN models yields relative confidence in their decision. To create the late fusion model, the wavelet responses were multiplied by CNN responses. The resulting maximum relative confidence of this multiplication is then used to predict outcomes.

3.5. Evaluation and comparison of models

Average accuracy was used to compare performance of models. This enabled comparability of the models with previous results, as average accuracy has been widely used [28,29]. Classification accuracy of each class was first determined by dividing the number of correct predictions over the total number of predictions. Average accuracy was then calculated by classification accuracy arithmetic mean of all classes. Confusion matrices were also presented for best models.

4. Results and discussion

4.1. Results of convolutional neural network (CNN) models

Fig. 4 shows the accuracy and loss during the training process for each iteration of the fifteen epochs included. The accuracy and loss are almost the same after epoch 8. The final overall average accuracy of CNN was 72.9% with SD ±20%. Fig. 5a shows the confusion matrix of CNN models for all included acoustic scenes.

4.2. Results of ensemble classifier models and their fusion with CNNs

Different ensemble classifiers were run as shown in Table 2. The average accuracy of ensemble classifier models ranged between 40.4 and 76.5%. The results of the fusion of these ensemble classifiers are also shown in Table 2. Random subspace had the highest average accuracy (76.5% with SD ±18%) and the highest average accuracy when lately fused with the CNN model (80%).

It was hypothesised that using late fusion models for ASC could yield more accurate models, compared to ordinary CNNs. The results in Table 2 confirmed this hypothesis. The confusion matrix for random subspace classifier and CNN-Random subspace classifier fusion model is shown in Fig. 5. There was a variability of the accuracy of classifying different acoustic scenes for the best fusion model (mean ± SD) 80 ± 17%.

4.3. Average accuracy comparison with previous studies

The results of the current study were compared with previous studies that used early fusion CNN models of the same dataset as

shown in Table 3. These results suggest that the late fusion of CNN with ensemble classifier could be a potential solution for future ASC problems as they could have higher average accuracy than early fusion models.

5. Conclusion

Accurate acoustic scene classification (ASC) models are of great help in many areas. This study presented an enhanced model for ASC by the late fusion of convolutional neural networks (CNNs) and ensemble classifiers. The results showed that the late fusion model had a higher accuracy for ASC, compared to the individual convolutional neural network (CNN) or ensemble classifier models. This fusion model had an average increase in accuracy of 10% as compared to the CNN model average accuracy. A comparison with previous studies using CNN models showed that the late fusion between CNN and ensemble classifier models can yield higher average accuracy, compared to early fusion CNN models (at least 7% increase). This suggests that the proposed late fusion could have promising applications for ASC problems.

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.

References

[1] Waldekar S, Saha G. Two-level fusion-based acoustic scene classification. Appl Acoust 2020;170:107502. https://doi.org/10.1016/j.apacoust.2020.107502.
[2] Mesaros A, Heittola T, Virtanen T. Acoustic scene classification: An overview of dcase 2017 challenge entries. In: 16th Int Work Acoust Signal Enhanc IWAENC 2018 - Proc. p. 411–5. https://doi.org/10.1109/IWAENC.2018.8521242.
[3] Bianco MJ, Gerstoft P, Traer J, Ozanich E, Roch MA, Gannot S, et al. Machine learning in acoustics: Theory and applications. J Acoust Soc Am 2019;146:3590–628. https://doi.org/10.1121/1.5133944.
[4] Sharan RV, Moir TJ. Acoustic event recognition using cochleagram image and convolutional neural networks. Appl Acoust 2019;148:62–6. https://doi.org/10.1016/j.apacoust.2018.12.006.
[5] Yang L, Tao L, Chen X, Gu X. Multi-scale semantic feature fusion and data augmentation for acoustic scene classification. Appl Acoust 2020;163:107238. https://doi.org/10.1016/j.apacoust.2020.107238.
[6] Su Y, Zhang K, Wang J, Zhou D, Madani K. Performance analysis of multiple aggregated acoustic features for environment sound classification. Appl Acoust 2020;158. https://doi.org/10.1016/j.apacoust.2019.107050.
[7] Zhang T, Liang J, Ding B. Acoustic scene classification using deep CNN with fine-resolution feature. Expert Syst Appl 2020;143:. https://doi.org/10.1016/j.eswa.2019.113067113067.
[8] Mulimani M, Koolagudi SG. Robust acoustic event classification using fusion fisher vector features. Appl Acoust 2019;155:130–8. https://doi.org/10.1016/j.apacoust.2019.05.020.
[9] Dong X, Yan Y, Tan M, Yang Y, Tsang IW. Late fusion via subspace search with consistency preservation. IEEE Trans Image Process 2019;28:518–28. https://doi.org/10.1109/TIP.2018.2867747.
[10] Li G, Ming Z, Li H, Chua TS. Early versus late fusion in semantic video analysis. In: Proc Seventeen ACM Int Conf Multimed. p. 773–6.
[11] Paseddula C, Gangashetty SV. Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks. Appl Acoust 2021;172:107568. https://doi.org/10.1016/j.apacoust.2020.107568.
[12] Atmaja BT, Akagi M. Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. IEEE 2020:4477–81.
[13] Pei E, Jiang D, Sahli H. An efficient model-level fusion approach for continuous affect recognition from audiovisual signals. Neurocomputing 2020;376:42–53. https://doi.org/10.1016/j.neucom.2019.09.037.
[14] Tsanousa A, Meditskos G, Vrochidis S, Kompatsiaris I. A weighted late fusion framework for recognizing human activity from wearable sensors. In: 10th Int Conf Information, Intell Syst Appl IISA 2019. https://doi.org/10.1109/IISA.2019.8900725.
[15] Mesaros A, Heittola T, Virtanen T. TUT Acoustic scenes 2017; 2017.
[16] Alamir MA, AlHares A, Hansen KL, Elamer A. The effect of age, gender and noise sensitivity on the liking of food in the presence of background noise. Food Qual Prefer 2020:84.
[17] Alamir MA, Hansen K. The effect of type and level of background noise on food liking: A laboratory non-focused listening test. Appl Acoust 2021;172:107600. https://doi.org/10.1016/j.apacoust.2020.107600.
[18] Alamir MA, Hansen KL, Zajamsek B, Catcheside P. Subjective responses to wind farm noise: A review of laboratory listening test methods. Renew Sustain Energy Rev 2019;114:. https://doi.org/10.1016/j.rser.2019.109317109317.
[19] Alamir MA, Hansen KL, Zajamsek B. The effect of wind farm noise on human response: An analysis of listening test methodologies. In: Proc. Acoust. 2018, Adelaide, Australia; 2018. p. 1–9.
[20] Alamir MA, Elamer AA. A compromise between the temperature difference and performance in a standing wave thermoacoustic refrigerator. Int J Ambient Energy 2018;0750:1–13. https://doi.org/10.1080/01430750.2018.1517673.
[21] Alamir MA. Experimental study of the stack geometric parameters effect on the resonance frequency of a standing wave thermoacoustic refrigerator. Int J Green Energy 2019.
[22] Alamir MA. Experimental study of the temperature variations in a standing wave loudspeaker driven thermoacoustic refrigerator. Therm Sci Eng Prog 2019:100361. https://doi.org/10.1016/j.tsep.2019.100361.
[23] Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Data-Dependent Data Augmentation; 2017.
[24] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proc IEEE 2016;104:148–75. https://doi.org/10.1109/JPROC.2015.2494218.
[25] Kulkarni P, Sadasivan J, Adiga A, Seelamantula CS. EPOCH Estimation from a speech signal using gammatone wavelets in a scattering network Department of Electrical Engineering , Indian Institute of Science , Bengaluru - 560012 , India Biocomplexity Institute and Initiative , University of Virginia , Charlo. ICASSP 2020 - 2020 IEEE Int Conf Acoust Speech Signal Process; 2020:7359–63.
[26] Alamir MA. An artificial neural network model for predicting the performance of thermoacoustic refrigerators. Int J Heat Mass Transf 2021;164:. https://doi.org/10.1016/j.ijheatmasstransfer.2020.120551120551.
[27] Alamir MA. Thermoacoustic energy conversion devices: novel insights. J Adv Res Fluid Mech Therm Sci 2021:77.
[28] Weiping Z, Jiantao Y, Xiaotao X, Xiangtao L, Shaohu P. Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion. Work Detect Classif Acoust Scenes Events 2017:1–5.
[29] Hyder R, Ghaffarzadegan S, Feng Z, Hasan T. Buet Bosch Consortium (B2C) acoustic scene classification systems for Dcase 2017 Challenge. DCASE 2017-Workshop Detect Classif Acoust Scenes Events, 2017.
[30] Lehner B, Eghbal-Zadeh H, Dorfer M, Korzeniowski F, Koutini K, Widmer G. Classifying short acoustic scenes with i-vectors and cnns: challenges and optimisations for the 2017 dcase asc task. DCASE 2017-Workshop Detect Classif Acoust Scenes Events, 2017.
[31] Park S, Mun S, Lee Y, Ko H. Acoustic scene classification based on convolutional neural network using double image features. Work Detect Classif Acoust Scenes Events 2017.
[32] Piczak KJ. The details that matter: frequency resolution of spectrograms in acoustic scene classification. DCASE 2017-Workshop Detect Classif Acoust Scenes Events, 2017.