An Ensembled Convolutional Recurrent Neural Network approach for Automated Classroom Sound Classification

Rashed Iqbal
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
NSW, Australia
mri510@uowmail.edu.au

Christian Ritz
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
NSW, Australia
critz@uow.edu.au

Jack Yang
School of Computing and Information Technology
University of Wollongong
NSW, Australia
jiey@uow.edu.au

Sarah Howard
School of Education
University of Wollongong
NSW, Australia
sahoward@uow.edu.au

Abigail Copiaco
College of Engineering and Information Technology
University of Dubai,
Dubai, UAE
acopiaco@ud.ac.ae

***Abstract*—The paper explores automated classification techniques for classroom sounds to capture diverse learning and teaching activities' sequences. Manual labeling of all recordings, especially for long durations like multiple lessons, poses practical challenges. This study investigates an automated approach employing scalogram acoustic features as input into the ensembled Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (BiGRU) hybridized with Extreme Gradient Boost (XGBoost) classifier for automatic classification of classroom sounds. The research involves analyzing real classroom recordings to identify distinct sound segments encompassing teacher's voice, student voices, babble noise, classroom noise, and silence. A sound event classifier utilizing scalogram features in an XGBoost framework is proposed. Comparative evaluations with various other machine learning and neural network methodologies demonstrate that the proposed hybrid model achieves the most accurate classification performance of 95.38%.**

Keywords—classroom activity, deep learning, sound classification, audio processing, artificial intelligence.

I. INTRODUCTION

Different types of sounds are present in a classroom environment, reflecting the diverse learning activities happening within. These encompass the teacher's speech, student queries, group interactions, and ambient noises. For researchers focusing on education, identifying patterns of these activities can lead to greater understanding of student learning in the classroom. Manually analyzing extensive recordings to find these learning activities over many days or weeks becomes challenging. Thus, this study delves into exploring an automated method centered on classifying sounds captured in classroom audio recordings.

Exploration in sound classification spans various fields like automated speech recognition, music genre differentiation, and ambient sound categorization [1]. To analyze classroom sounds, the focus is on recording methods that least disrupt the natural learning process. Typically, this involves situating a microphone in a single position within the classroom rather than attaching individual close-talking microphones (e.g., lapel mics) to each student and the teacher. Previously, the authors devised a system for capturing classroom video and audio to investigate technology-driven learning [2]. These audio recordings, obtained from a singular source near the teacher, encompass the teacher's speech, student dialogues, and incidental sounds within the classroom. Challenges arise due to background noise and varying distances between the fixed microphone and moving students or teachers during lessons, impacting the quality of these recordings. Consequently, automatically classifying such audio presents more complexity compared to other sound classification endeavours.

Numerous deep learning approaches [3-7] have surfaced for the supervised automated sound classification task. A recent development in the DCASE 2022 Task 1 introduces an extensive dataset crafted [8] for lower complex sound classification, offering segment wise labelled data. Methods centered on various device data partition into segments and associate simulated data during the training phase. Presently, there exists a research gap in evaluating and contrasting limited weakly labelled training data preparation and segment wise classification methodologies tailored for sound classification. While convolutional neural networks (CNN)-based methods have shown proficiency in audio tagging and classification tasks, their limitation lies in effectively capturing temporal relationships within an audio clip. Techniques like CNNs as highlighted in [6, 9, 10] have been employed to address this by incorporating prolonged temporal information for supervised sound classification. However, one drawback of CNNs is their sequential computation of hidden states, lacking parallel processing capabilities.

In earlier studies, the author investigated using sound power level attributes derived from the Decibel Analysis for Research in Teaching (DART) algorithm [11] to categorize classroom audio obtained through their developed system [2]. Other researchers have also delved into various methods for

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00041

analyzing classroom audio. For instance, a Multi-Scale Audio Spectrogram Transformer (MAST) was created to identify interactions between teachers and students during classroom sessions [12]. However, this research primarily concentrated on verbal exchanges between teachers and students, overlooking other potential classes crucial for comprehending key learning activities.

Current methods of categorizing classroom sounds [13, 14] involve training neural networks with features extracted from labelled classroom sound recordings. The DART technique relies on uncomplicated features derived from sound power levels, akin to the approaches employed in [13]'s neural networks. Conversely, time-frequency attributes obtained from the audio recordings, like the mel-spectrogram [14], are commonly utilized. Neural networks utilized for classifying classroom audio in [13, 14] encompass Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and RNN variations like Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. This classification aligns with the broader domain of environmental sound or scene classification (for an elaborate review, refer to [15, 16]).

This study introduces scalogram features derived from the Continuous Wavelet Transform (CWT) as an innovative approach for classifying classroom sounds. These features, outperform traditional time-frequency-based features such as mel-spectrum features in audio classification using neural networks [17-19], present a promising alternative. An obstacle in using classroom audio recordings for classification lies in the scarcity of labelled training data, given the authentic classroom settings. Training a single neural network model under these circumstances poses challenges in achieving optimal performance. Hence, this paper explores hybrid methods that amalgamate CNN and Bidirectional Gated Recurrent Unit (BiGRU), as BiGRU performs better than LSTM and BiLSTM in terms of categorical classification [20] networks to generate features, subsequently used an Extreme Gradient Boost (XGBoost) classifier. Which is motivated from the CNN-XGBoost combination having highest classification accuracy of 99.22% in [21]. These techniques aim to automatically categorize the previously mentioned five sound classes, deviating from prior studies focusing on fewer categories.

Section II outlines the classroom sound recording process, signal pre-processing, feature extraction, and the proposed classification systems. Section III encompasses an in-depth analysis of the experimental evaluation, including the optimization steps for the hybrid model and the performance comparisons across different models. The paper concludes with key insights and suggestions for future research.

II. Classroom audio recordings and preprocessing

This section provides an overview of the classroom audio recording procedure and preprocessing.

A. Classroom audio recordings

The sound in a ninth-grade science classroom in a city-based Australian high school was recorded. This research, building upon a previous project, was approved by the University of Sydney's ethics board and endorsed by the New South Wales (NSW) Department of Education. Permission was granted from the students, their parents, and the instructor. The classroom implemented a Bring Your Own Device (BYOD) strategy, utilizing Microsoft OneNote for studying and note-taking. It was furnished with four cameras and audio recorders. There were 25 students and one teacher in the class, meeting for 80 minutes four times weekly. Recording devices were only activated during the scheduled Year 9 sessions.

TABLE I. Preparation of classroom data

Training and validation set (1st and 2nd hour)			**Evaluation set (3rd hour)**
Class ID	*Classes*	*Number of segments*	
1	Teacher	248	62
2	Student	216	54
3	Classroom Sound	100	25
4	Babble Noise	256	64
5	Silence	100	25

B. Data pre-processing

The training dataset for the model was derived from a specific two hours segment of genuine classroom recordings. This segment underwent manual scrutiny and classification, resulting in the creation of a database with 1150 five second audio clips categorized into the five classes outlined in Table I. These classes include "teacher," denoting moments when the teacher delivers lectures or instructions, "student," indicating instances where a student asks or responds to questions, "classroom sound," is a sound that occurs typically in classroom environment such as Shoes sound, experimental instrument sounds, some stroking sound on the bench, laughing, shouting loud, "babble noise," corresponding mainly to the beginning or end of the class, before or after the lesson, and "silence," primarily encompassing periods when students engage in silent learning activities as directed by the teacher. The audio files, sampled at 32 kHz, were initially stored as MP3 files compressed at 96 kbps; later, they were down sampled to 16 kHz for integration into the classification system. The next one hour of the data was used for evaluation, which includes 230 five second slices.

Scalograms are created and expanded to prepare for training and testing deep learning models. To properly assess these models, the dataset is cross validated. Yet, it's important to highlight that the sample distribution among classes isn't consistent, as depicted in Table I.

III. Classroom Audio Classification System

The sound classification system illustrated in Figure 1 extracts scalogram characteristics from the audio input to fuel an ensembled model that combines CNN-BiGRU-XGBoost components. This study explores and contrasts the performance of pre-trained models such as ResNet50, VGG16, MobileNetV2, and InceptionV3 in the context of transfer learning. As transfer learning has demonstrated remarkable accuracy in environmental sound classification [22] and other commonly used CRNN models like CNN-LSTM, CNN-PSO, CNN-BiLSTM, and CNN-BiGRU.

A. Scalogram conversion

Unlike the fixed resolution of the Short Time Fourier Transform (STFT), the Continuous Wavelet Transform

(CWT) (1) operates using adjusted versions of a base wavelet. The scalogram, akin to a spectrogram, depicts the absolute values of the CWT coefficients across time and scale in a two-dimensional format [23]. This representation has demonstrated superior efficacy compared to other time-frequency characteristics employed in neural network-driven audio classification [17].

$$CWT_c(s,t) = \int_{-\infty}^{\infty} x(u)\frac{1}{\sqrt{s}}\psi^*\left(\frac{u-t}{s}\right)du \qquad (1)$$

Within this framework (1), $x(u)$ is the input signal, s represents the scale parameter which is contingent on the frequency, ψ^* represents the conjugate of the main wavelet, t is the translation parameter, which shifts the wavelet function along the time axis, and u stands for the signal segment.

B. *Ensembled CNN-BiGRU model*

The Ensemble CNN-BiGRU model represents a powerful architecture merging Convolutional Neural Network (CNN) layers with Bidirectional Gated Recurrent Unit (BiGRU) layers. The model is designed to handle sequential data, assuming an input format of 4D tensors. The model is compiled using sparse categorical cross entropy as the loss function. This ensemble strategy involves training multiple instances of this architecture using Stratified K-Fold Cross-Validation, storing the trained models in an ensemble. The ensemble aims to leverage the spatial feature extraction ability of CNNs and the temporal dependency capturing capability of BiGRUs to enhance predictive accuracy and robustness in identifying intricate patterns within sequential data. The CNN (2) extracts feature from the scalogram. In each convolution layer index by l, convolution operation and an additive bias will be applied to the input, for a feature map indexed by $f \in \{1, \dots., f(l)\}$. The output $y_i^{(l)}$, of the l^{th} layer for the i^{th} feature map, is derived from previous layer $y_i^{(l-1)}$ [24].

$$y_i^{(l)} = \emptyset(B_i^{(l)} + \sum_{j=1}^{f(l-1)} K_{i,j}^{(l)} * y_j^{(l-1)}) \qquad (2)$$

Where $\emptyset$ is rectified linear unit (ReLU), $B_i^{(l)}$ is a bias matrix, $K_{i,j}^{(l)}$ is filter size.

The BiGRU processes the features from the CNN across time steps in (5), through forward (3) and backward (4) states.

$$\overrightarrow{h_t} = GRU_{fwd}(h_t, \vec{h}_{t+1}) \qquad (3)$$

$$\overleftarrow{h_t} = GRU_{bwd}(h_t, \vec{h}_{t+1}) \qquad (4)$$

$$\bar{h}_t = \overrightarrow{h_t} + \overleftarrow{h_t} \qquad (5)$$

$\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ represents the forward and backward hidden states generated by the BiGRU layer from the feature map.

C. *K-Fold Cross-Validation*

The model employs five Stratified K-Fold Cross-Validation using the Stratified K-Fold method to create stratified folds for robust model assessment, effectively splitting the dataset while considering associated labels. Employing an enumeration loop through the generated folds using enumerate skf.split, the data is systematically divided into training and testing sets based on fold indices. This segmentation operation results in the creation of distinct subsets, where x_train and x_test represent the training and testing data splits respectively, while y_train and y_test contain the corresponding labels for training and testing. This process facilitates the evaluation of model performance across different subsets of the data, ensuring a comprehensive understanding of the model's behaviour and generalization capabilities across various segments of the dataset. The whole dataset was divided into two sets, first tow hours of the recording annotated for training and cross validation.

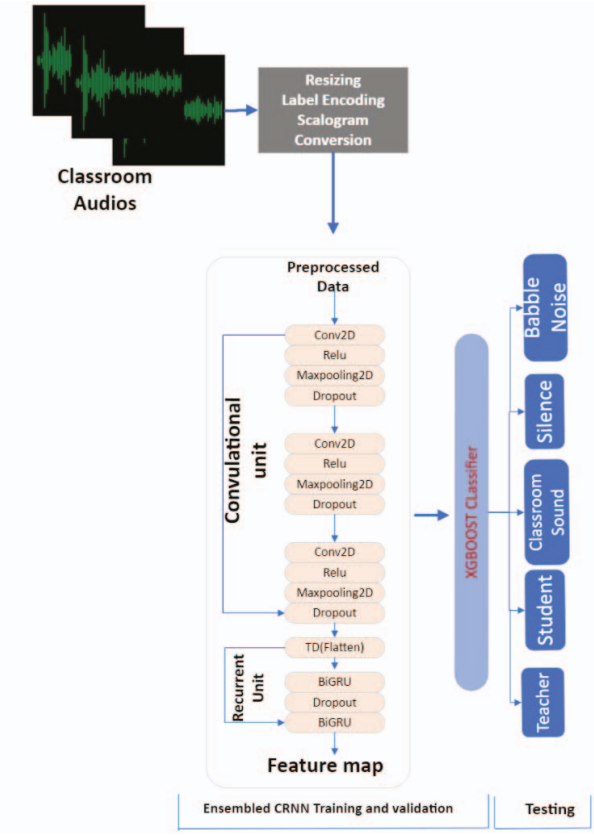

Fig.1. Process flow diagram of the proposed model

D. *Integration of XGBoost*

The integration of XGBoost models atop the CNN-BiGRU ensemble aims to further refine predictions and enhance model performance. The XGBoost model is utilized to make predictions on the test data obtained from the CNN-BiGRU model. This ensemble strategy leverages the unique strengths of both CNN-BiGRU and XGBoost models to improve overall predictive accuracy and generalize well on unseen data [25]. It uses an ensemble of K classification and regression trees. Each of which has $K_E^i | i \in 1 \dots K$ nodes. The final prediction scores for each tree in 6.

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{6}$$

Where the x_i are the members of the training set and y_i are the corresponding class labels, f_k is the leaf score for the k^{th} tree and F is the set of all K scores for all classification and regression trees. Regularization is applied to improve the final result in (7).

$$L(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{7}$$

The first term l, represents the differentiable loss function, which measures the difference between target y_i and prediction $\hat{y}_i$. The second term in (8) to avoid overfitting:

$$\Omega(f_k) = \gamma^T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

Where γ, λ are constants controlling the regularization degree, T is the number of leaves in the tree and w is the weight of each leaf. Gradient boosting is effective in regression classification problems.

E. Evaluation metrics

Key metrics like Accuracy, Precision, Recall, and F1 Score are crucial in evaluating classification model performance. Accuracy (ACC) (9) measures overall correctness, while Precision (10) assesses the accuracy of positive predictions, and Recall (11) evaluates the model's ability to identify all actual positive instances. The F1 Score (12), a blend of Precision and Recall, is particularly useful for imbalanced class distributions as it balances these metrics. While Accuracy provides a general view, Precision and Recall offer specific insights. The F1 Score, considering false positives and false negatives, provides a consolidated metric for a balanced assessment of Precision and Recall [26]. Together, these metrics help thoroughly assess a model's predictive capabilities, revealing its strengths and weaknesses in classification tasks.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$Precesion = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{Precesion\,.\,Recall}{Precesion + Recall} \tag{12}$$

IV. Experimental Results and Analysis

In this segment, outcomes from the training and testing phases are showcased, employing evaluation criteria including accuracy, precision, recall, and F1 scores. Additionally, diverse neural network models, encompassing transfer learning and different CRNN variations, were utilized for comparative analysis and assessment.

A. Experiment setup

Pretrained models are retrained and tested with our own dataset where the categorical cross-entropy loss function was used. It measures the difference between the predicted probability distribution and the actual distribution of the classes. This function penalizes the model more significantly for larger deviations from the correct class, encouraging the model to minimize the error in predicting the correct class probabilities.

A sequential input layer started with three sets of Conv2D layers with increasing filters (32, 64, and 128) and ReLU activation. MaxPooling2D layers with a pool size of (2, 2) after each Conv2D layer to reduce spatial dimensions. Dropout layers with a dropout rate of 0.25 after each MaxPooling2D layer to prevent overfitting. A Time Distributed (Flatten ()) employed to flatten the data before feeding it into the recurrent layer, two bidirectional GRU layers with 64 and 32 units, respectively are added. Post-training, the data is reshaped for integration with an XGBoost classifier, which is fitted with the training data. The same loss function is also used in our proposed model as pretrained models. The code combines predictions from multiple ensemble stages, averaging their outputs to produce final predictions, evaluated for accuracy against test data in each fold.

B. Training and validation results

The training results of Table II showcase the performance of various models across accuracy, validation accuracy, training loss, validation loss, and parameter counts. Among these models, CNN-BiGRU-XGBoost achieved the highest accuracy of 97.58% and a validation accuracy of 96.68%, demonstrating its superior predictive capability. It showcased the lowest training loss of 0.07 and validation loss of 0.10, indicative of its ability to generalize well on unseen data. Notably, Inception V3 and CNN-BiGRU also performed admirably with accuracies of 93.65% and 92.97%, respectively, showcasing strong classification capabilities. MobileNetV2 exhibited high accuracy at 91.52%, emphasizing its efficiency with significantly fewer parameters (3.5 million). These results indicate the diverse strengths and capabilities of each model, from high accuracy with lower parameter counts (MobileNetV2) to more complex architectures (Inception V3, CNN-BiGRU-XGBoost) achieving remarkable accuracy and generalization.

TABLE II. Training Performance of the Models

Models	Accuracy	Validation	Training loss	Validation loss	Param
ResNet50	40.86%	41.66%	1.30	1.29	25 M
VGG16	68.40%	68.60%	0.78	0.79	138 M
MobileNet V2	91.52%	90.47%	0.26	0.29	3.5 M
Inception V3	93.65%	91.19%	0.19	0.27	23 M
CNN-LSTM	88.0%	92.33%	0.37	0.23	0.31 M
CNN-PSO	81.72%	84.35%	0.48	0.38	0.31 M
CNN-BiLSTM	85.03%	84.84%	0.42	0.32	0.72 M
CNN-BiGRU	92.97%	91.50%	0.21	0.28	0.24 M
CNN-BiGRU-XGBoost	97.58%	96.68%	0.07	0.10	0.24 M

The Stratified K-Fold Cross-Validation of the CRNN-XGBoost model in Table III, demonstrates consistent and

robust performance across five folds. Each fold's accuracy ranged from 96.1% to 98.9%, with an average accuracy of 97.58%. Precision scores remained high, consistently above 0.97, indicating the model's ability to correctly identify positive cases among the predicted ones. Similarly, recall scores were consistently strong, hovering around 0.98 in most folds, showcasing the model's proficiency in capturing all positive instances. The F1 scores, reflecting the harmonic mean of precision and recall, maintained a commendable performance, averaging 0.970 across all folds. These results collectively signify the model's stability, reliability, and effectiveness in accurately classifying data while maintaining a balance between precision and recall.

TABLE III. STRATIFIED K-FOLD CROSS VALIDATION SUMMARY

Folds	ACC	Precision	Recall	F1
1	96.10%	0.965	0.961	0.959
2	97.30%	0.974	0.976	0.960
3	98.90%	0.982	0.979	0.982
4	97.50%	0.975	0.977	0.970
5	98.10%	0.980	0.981	0.979
Avg	97.58%	0.975	0.974	0.970

C. Evaluation of the models

In Table IV, The CNN-BiGRU-XGBoost model exhibited exceptional performance when evaluated against unseen testing data. With an impressive accuracy of 95.38%, it outperformed various other models in the comparative analysis. It showcased robustness and precision with a high F1 score of 0.951, indicating a strong balance between precision and recall. The model's precision of 0.943 signifies its ability to correctly identify positive cases, while the recall of 0.953 demonstrates its capability to accurately capture all positive instances within the dataset. Overall, the CNN-BiGRU-XGBoost model demonstrated superior predictive ability, highlighting its effectiveness in handling real-time predictions for the next hour's data.

TABLE IV. PERFORMANCE EVALUATION THE TESTING DATA

Models	Accuracy	Precision	Recall	F1
ResNet50	28.26%	0.421	0.28	0.210
VGG16	52.60%	0.423	0.526	0.456
MobileNetV2	81.73%	0.851	0.817	0.807
Inception V3	92.17%	0.945	0.921	0.925
CNN-LSTM	80.0%	0.753	0.80	0.782
CNN-PSO	71.32%	0.684	0.713	0.705
CNN-BiLSTM	80.21%	0.848	0.802	0.80
CNN-BiGRU	90.70%	0.925	0.907	0.902
CNN- BiGRU-XGBoost	95.38%	0.943	0.953	0.951

The confusion matrix in Fig.2 outlines the performance of the CRNN-XGBoost model on the testing data. Each row of the matrix represents the true labels, while each column depicts the predicted labels. For instance, in the "Babble Noise" class, the model correctly predicted 52 instances, misclassifying 1 as "Classroom Sound," 9 as "Silence," and 2 as "Student." Moreover, no instances from "Babble Noise" were falsely predicted as "Teacher." Similarly, for the "Teacher" class, the model accurately predicted all 62 instances without misclassification. The confusion matrix provides a comprehensive view of other model's performance across different classes, revealing where the model excels and areas where it may have challenges in accurately predicting specific labels.

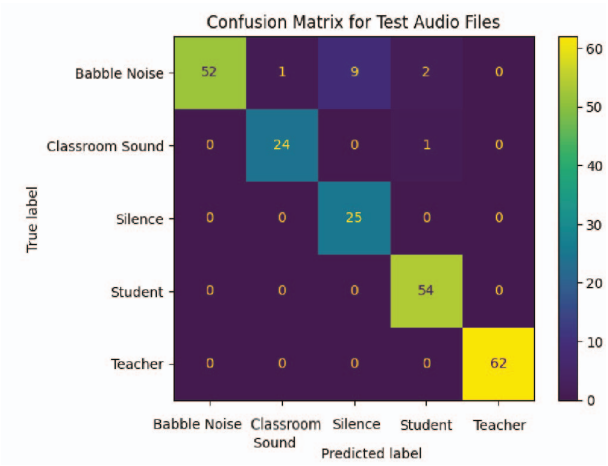

Fig.2. Confusion matrix of the testing data

Overall, the proposed model demonstrates the highest capabilities of classification across all classes in real-time identification. Which can be seen from the above confusion matrix.

Moreover, in Fig.3 Our sound classification model displayed remarkable accuracy across diverse datasets, achieving 95.38% on our own Classroom data, 96.25% on ESC10, and 92.51% on ESC50. Notably, it excelled further on the Urban Sound dataset, attaining an impressive accuracy of 97.87%. These consistent high accuracies validate its robustness and generalization ability across varied sound environments.

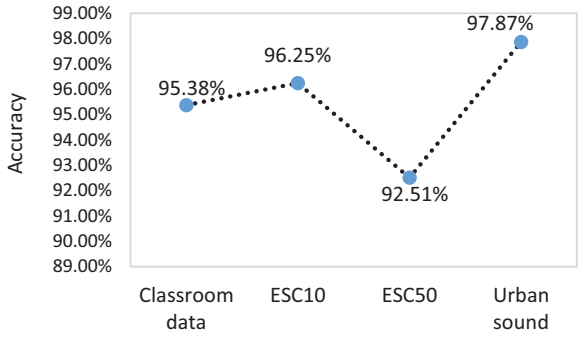

Fig.3. Accuracy comparison across diverse dataset

V. CONCLUSION

The proposed ensemble learning strategy combines CNN models with BiGRU layers and employs K-fold cross-validation for training and validation. Each fold involves constructing and training CRNN models, using their predictions on test data as features for subsequent XGBoost classifiers. The use of scalogram features for audio representation is a significant advantage, contributing to

effective feature extraction in the CRNN model. This ensemble approach synergizes the strengths of CNNs, RNNs, and gradient boosting, enhancing predictive performance across the dataset. Future work will involve testing the classification model with different audio spectra, especially in real-time classroom environments in Australia, with a focus on incorporating larger training datasets for improved accuracy and generalization.

ACKNOWLEDGMENT

This research data was obtained from the ARC Discovery Project DP130100481.

REFERENCES

[1] E. Tsalera, A. Papadakis, and M. Samarakou, "Novel principal component analysis – based feature selection mechanism for classroom sound classification," *Computational Intelligence,* vol. 37, no. 4, pp. 1827-1843, 2021.

[2] S. K. Howard, J. Yang, J. Ma, C. Ritz, J. Zhao, and K. Wynne, "Using data mining and machine learning approaches to observe technology-enhanced learning," in *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 2018: IEEE, pp. 788-793.

[3] G. Petmezas *et al.*, "Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function," *Sensors,* vol. 22, no. 3, p. 1232, 2022.

[4] L. Zhang, C. P. Lim, Y. Yu, and M. Jiang, "Sound classification using evolving ensemble models and Particle Swarm Optimization," *Applied soft computing,* vol. 116, p. 108322, 2022.

[5] S. D. H. Permana, G. Saputra, B. Arifitama, W. Caesarendra, and R. Rahim, "Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm," *Journal of King Saud University-Computer and Information Sciences,* vol. 34, no. 7, pp. 4345-4357, 2022.

[6] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep CNN model for environmental sound classification," *IEEE Access,* vol. 8, pp. 66529-66537, 2020.

[7] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters,* vol. 24, no. 3, pp. 279-283, 2017.

[8] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623,* 2020.

[9] J. Xie, K. Hu, M. Zhu, J. Yu, and Q. Zhu, "Investigation of different CNN-based models for improved bird sound classification," *IEEE Access,* vol. 7, pp. 175353-175361, 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, p. 107520, 2020.

[11] M. T. Owens *et al.*, "Classroom sound can be used to classify teaching practices in college science courses," *Proceedings of the National Academy of Sciences,* vol. 114, no. 12, pp. 3085-3090, 2017.

[12] F. Liu and J. Fang, "Multi-Scale Audio Spectrogram Transformer for Classroom Teaching Interaction Recognition," *Future Internet,* vol. 15, no. 2, p. 65, 2023.

[13] R. Cosbey, A. Wusterbarth, and B. Hutchinson, "Deep learning for classroom activity detection from audio," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: IEEE, pp. 3727-3731.

[14] A. Mou, M. Milanova, and M. Baillie, "Active Learning Monitoring in Classroom Using Deep Learning Frameworks," in *International Conference on Pattern Recognition*, 2022: Springer, pp. 384-393.

[15] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing,* vol. 13, no. 2, pp. 206-219, 2019.

[16] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences,* vol. 10, no. 6, p. 2020, 2020.

[17] A. Copiaco, C. Ritz, N. Abdulaziz, and S. Fasciani, "A study of features and deep neural network architectures and hyper-parameters for domestic audio classification," *Applied Sciences,* vol. 11, no. 11, p. 4880, 2021.

[18] H. Pham Thi Viet, H. Nguyen Thi Ngoc, V. Tran Anh, and H. Hoang Quang, "Classification of lung sounds using scalogram representation of sound segments and convolutional neural network," *Journal of Medical Engineering & Technology,* vol. 46, no. 4, pp. 270-279, 2022.

[19] Ö. İnik, "CNN hyper-parameter optimization for environmental sound classification," *Applied Acoustics,* vol. 202, p. 109168, 2023.

[20] L. Zhou and X. Bian, "Improved text sentiment classification method based on BiGRU-Attention," in *journal of physics: conference series*, 2019, vol. 1345, no. 3: IOP Publishing, p. 032097.

[21] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, "A novel image classification method with CNN-XGBoost model," in *Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017, Proceedings 16*, 2017: Springer, pp. 378-390.

[22] A. Ashurov, Y. Zhou, L. Shi, Y. Zhao, and H. Liu, "Environmental Sound Classification Based on Transfer-Learning Techniques with Multiple Optimizers," *Electronics,* vol. 11, no. 15, p. 2279, 2022.

[23] J. Lu, R. Ma, G. Liu, and Z. Qin, "Deep convolutional neural network with transfer learning for environmental sound classification," in *2021 International Conference on Computer, Control and Robotics (ICCCR)*, 2021: IEEE, pp. 242-245.

[24] S. Thongsuwan, S. Jaiyen, A. Padcharoen, and P. Agarwal, "ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost," *Nuclear Engineering and Technology,* vol. 53, no. 2, pp. 522-531, 2021.

[25] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.

[26] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint arXiv:2008.05756,* 2020.