

Scene Classification via Attention-Guided Integration of Visual and Auditory Data Streams

Yusuf ÇELİK^{1*} 

¹Munzur University, Engineering Faculty, Department of Computer Engineering, Tunceli, Türkiye

**Corresponding author: celikyusuf@munzur.edu.tr*

Abstract

This study proposes a multi-source deep learning architecture, named the Gated Cross-Modal Attention Fusion Network (GCM-AFN), for integrating visual and auditory information in scene classification tasks. The proposed framework extracts deep features from the visual stream using an EfficientNetV2 backbone, while MFCC-based time-frequency representations are used for the auditory stream. The feature representations obtained from both streams are fused through a gated attention mechanism. The model incorporates auxiliary stream outputs and an attention-based fusion strategy to effectively combine multimodal information. Experimental results based on cross-validation demonstrate that the proposed approach achieves higher accuracy, lower variance, and more consistent class-wise performance compared to single-stream models and conventional fusion methods. These findings suggest that attention-based fusion provides an effective integration strategy for visual-auditory scene classification.

Keywords

Audio-Visual
Scene
Classification,
Multimodal Deep
Learning,
Gated Attention
Fusion

Dikkat Tabanlı Görsel-İşitsel Akış Füzyonu ile Sahne Sınıflandırma

Yusuf ÇELİK^{1*} 

¹Munzur Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Tunceli, Türkiye

*Sorumlu yazar: celikyusuf@munzur.edu.tr

Özet

Bu çalışmada, sahne sınıflandırma görevlerinde görsel ve işitsel verilerin birlikte kullanılmasına yönelik, Gated Cross-Modal Attention Fusion Network (GCM-AFN) adı verilen çok modlu bir derin öğrenme mimarisi önerilmektedir. Önerilen yapıda, görsel akıştan EfficientNetV2 omurgası kullanılarak derin özellikler çıkarılırken, işitsel akışta MFCC tabanlı zaman-frekans temsilleri kullanılmaktadır. Her iki akıştan elde edilen özellik temsilleri, kapılı dikkat mekanizması aracılığıyla birleştirilmektedir. Model, çok kipli bilgilerin etkin biçimde bir araya getirilmesini sağlamak amacıyla yardımcı akış çıktıları ve dikkat tabanlı bir füzyon stratejisi içermektedir.

Çapraz doğrulamaya dayalı deneysel sonuçlar, önerilen yaklaşımın tek akışlı modeller ve geleneksel füzyon yöntemlerine kıyasla daha yüksek doğruluk, daha düşük varyans ve sınıf bazında daha tutarlı bir performans sergilediğini göstermektedir. Elde edilen bulgular, dikkat tabanlı füzyonun görsel-ışitsel sahne sınıflandırma görevlerinde etkili bir bilgi birleştirme yaklaşımı sunduğunu ortaya koymaktadır.

Anahtar kelimeler

Görsel-İşitsel Sahne Sınıflandırma, Çok Modlu Derin Öğrenme, Kapılı Dikkat Tabanlı Füzyon

1. INTRODUCTION

Artificial intelligence and deep learning techniques have become fundamental approaches in numerous fields, including speech recognition [1], image recognition [2,3], medical image classification tasks such as skin disease recognition [4] and COVID-19 detection from CT images [5], natural language processing [6], signal analysis [7], and scene classification using aerial imagery [8]. The increased availability of large-scale datasets and advancements in computational power have enabled models capable of autonomously learning complex patterns, leading to significant improvements in classification, detection, segmentation, and prediction tasks [9]. Despite these advances, several challenges persist in real-world settings such as noisy data, distributional discrepancies, limited labeled samples, hardware constraints, and the need for reliability and interpretability. Consequently, approaches relying on a single information type have become insufficient, prompting a shift toward integrated learning frameworks that combine multiple sources of information [10].

Scene classification is a computer vision task that aims to identify the environment represented in an image or video. Although early research predominantly focused on visual inputs, incorporating auditory cues has proven beneficial for enhancing contextual understanding. Sounds such as ocean waves in a beach scene or crowd noise during a football match can significantly strengthen scene recognition performance. However, effectively integrating information from multiple heterogeneous sources remains challenging due to differences in structure, distribution, noise levels, and completeness between audio and visual signals. Therefore, models designed for scene analysis must not only possess strong feature extraction capabilities but also incorporate flexible fusion mechanisms to balance interactions between distinct information sources.

A foundational study by Bird et al. [11] introduced a late-fusion deep learning approach in which features extracted separately from audio and visual data were merged at the decision stage. Their method, evaluated on the Look and Listen dataset comprising synchronized one-second audio–image samples across nine scene categories, employed a VGG16-based visual model and an MLP-based audio model. The system achieved approximately 96% accuracy, demonstrating the complementary nature of visual and auditory cues.

Following this work, larger-scale benchmark studies have been conducted. In the DCASE 2021 Challenge, the Audio-Visual Scene Classification task employed the TAU Audio-Visual Urban Scenes 2021 dataset, which includes recordings collected from ten cities across ten scene categories and spans roughly thirty-four hours of data. Analysis of the 43 submitted systems showed that the top-performing models consistently integrated audio and visual information, frequently leveraging pretrained architectures such as ResNet, EfficientNet, and Swin-Transformer through transfer learning. These systems employed combinations of early- and late-fusion

strategies and achieved accuracy levels between 93% and 94%, highlighting the superiority of joint audio–visual modeling over unimodal approaches [12].

Research following the challenge has expanded in methodological diversity. For example, Lei and Chen [13] introduced a multi-graph fusion network that models structural similarities between audio and visual samples, improving accuracy by 3–4% over previous fusion methods. Similarly, Chen et al. [14] proposed a joint optimization strategy that learns audio and visual representations simultaneously throughout training, resulting in more stable and robust performance. Pham et al. [15] investigated classification in crowded environments using audio–visual data derived from YouTube videos and demonstrated that fusion approaches outperform unimodal baselines in noisy conditions.

Attention-based architectures have also gained prominence in audio–visual scene analysis. Zhou et al. [16] presented a hierarchical attention network that integrates multiscale features and achieved improved performance on both the DCASE and AVE datasets, particularly for low-energy audio components. Xu et al. [17] extended multimodal learning concepts to industrial sensor data, showing that complementary information across different sources enhances predictive performance.

Other studies have explored feature selection and relational modeling. Liang et al. [18] introduced an association-based fusion method that captures statistical dependencies across data sources to increase system stability. The A3CarScene dataset proposed by Cantarini et al. [19] enabled audio–visual scene understanding in in-vehicle environments, supporting applications in autonomous driving. More recently, Sumathi et al. [20] developed a feature-selection-based scene recognition framework that eliminates redundant information and improves computational efficiency.

2. MATERIAL AND METHOD

2.1. Dataset

This study employs the Scene Classification – Images and Audio dataset, which follows the data generation protocol introduced by Bird et al. [11]. In the original study, the dataset was reported to contain 16,000 synchronized image–audio samples across eight environmental scene classes.

In the present work, we use the open-access version of the same dataset published on the Kaggle platform [21]. The Kaggle release expands the dataset, providing a total of 17,252 samples across nine scene categories. The dataset consists of synchronized RGB images and pre-extracted audio features represented as Mel-Frequency Cepstral Coefficients (MFCCs). All images were resized to 128×128×3, while audio signals were processed using overlapping 0.25-second windows to compute 13 MFCC coefficients, resulting in 104 time–frequency features per second. The class distribution of the Kaggle version used in this study is presented in Table 1.

Table 1. Class distribution and source counts of the Scene Classification – Images and Audio dataset used in the Kaggle version.

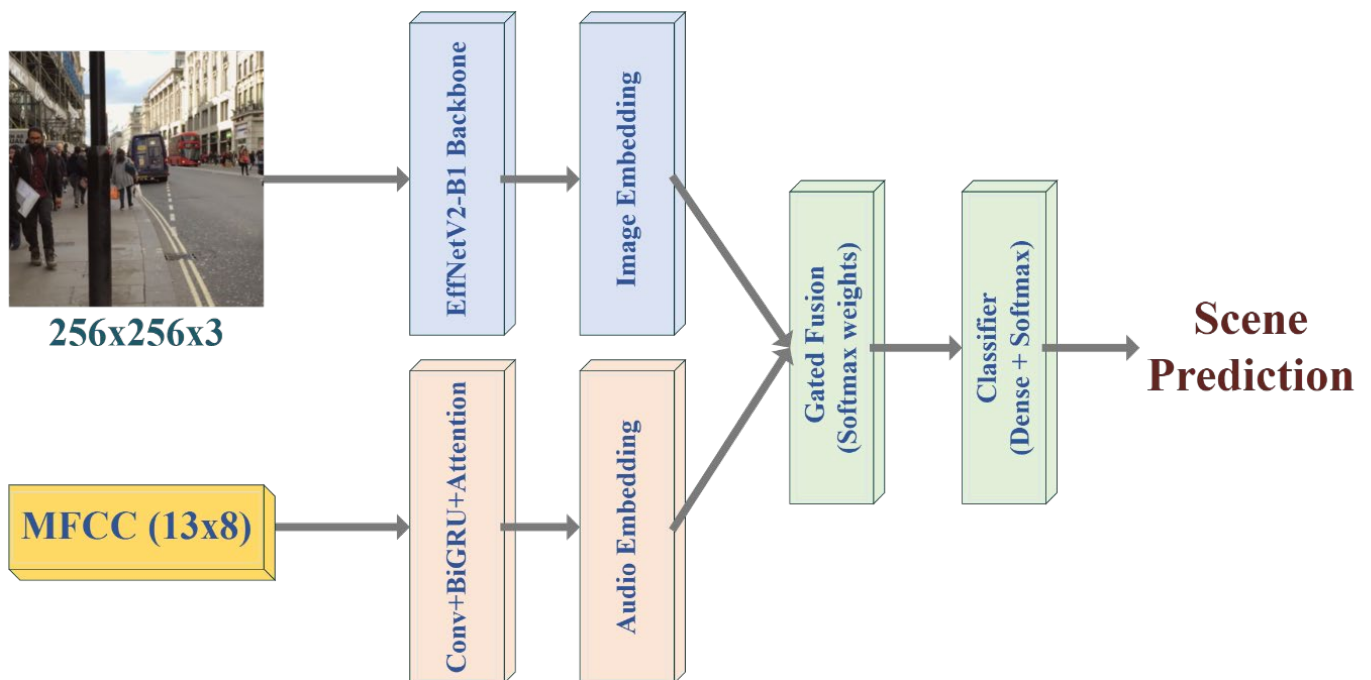
Class	Number of Samples	Percentage (%)	Source (Video)
Classroom	2753	15.96	6
City	2432	14.10	7
River	2084	12.08	8
Beach	2080	12.06	4
Grocery-store	2079	12.05	4
Football-match	1600	9.27	4
Restaurant	1459	8.46	8
Jungle	1404	8.14	2
Forest	1361	7.89	2
Total	17252	100 %	45

As shown in Table 1, the dataset exhibits a generally balanced class distribution. The classes with the highest number of samples are Classroom (2,753) and City (2,432), while the lowest number of samples appears in the Forest category (1,361). This diversity is important for evaluating the model's ability to generalize across both indoor and outdoor environments. Example images from the nine scene categories in the dataset are presented in Figure 1.

**Figure 1.** Example images from each class of the Scene Classification – Images and Audio dataset.

2.2. Proposed Model

In this study, a multimodal deep learning architecture called the Gated Cross-Modal Attention Fusion Network (GCM-AFN) is proposed for the scene classification problem. The overall structure of the model is illustrated in Figure 2. The system consists of two parallel processing streams that handle visual and auditory inputs, respectively. Feature representations extracted from these streams are combined through a gated attention-based fusion mechanism, and the final scene prediction is produced by a classifier layer.

**Figure 2.** General structure of the proposed GCM-AFN architecture. Features extracted from the visual and auditory streams are combined in a gated attention-based fusion block to generate the final class prediction.

The visual stream is configured using an EfficientNetV2-B1 backbone operating on RGB images from each scene. The network is initialized with ImageNet-pretrained weights, and only the final layers are fine-tuned during training. The input resolution is $256 \times 256 \times 3$. Data augmentation techniques such as rescaling, random cropping, horizontal flipping, and brightness/contrast adjustment are applied during training. The backbone output is reduced to 256 channels through a 1×1 convolution, followed by global average pooling to obtain a 256-dimensional visual feature vector.

The auditory stream processes Mel-Frequency Cepstral Coefficient (MFCC) features extracted from the audio samples corresponding to the same scene. Each MFCC tensor consists of 8 time steps and 13 coefficients. This stream applies Conv2D layers (64 and 128 filters), Batch Normalization, MaxPooling2D, a bidirectional GRU with 128 units, and an attention-based pooling layer in sequence. As a result, a 256-dimensional auditory feature representation is obtained for each audio sample.

The representations from the two streams are fused in the Gated Fusion block. At this stage, the visual and auditory feature vectors are combined, and soft weighting coefficients for each stream are dynamically learned through a Softmax layer. This allows the model to automatically determine which source of information is more dominant for each scene type. Through this mechanism, the architecture adaptively balances the contribution of each stream and simultaneously benefits from the complementary strengths of visual and auditory signals.

The fused feature vector is passed through a fully connected layer with 256 units (GELU activation) and a dropout rate of 0.2 before being fed into the classifier. The output layer consists of a fully connected structure with as many neurons as the number of scene classes, followed by a Softmax activation.

The training procedure follows a 10-fold shuffled stratified k-fold cross-validation protocol. A 10-fold stratified cross-validation strategy was adopted to ensure robust performance estimation and to reduce variance caused by data partitioning. For each fold, training is performed for 100 epochs using the AdamW optimizer. The learning rate is linearly increased during the first five epochs using a warm-up strategy and then decayed following a cosine schedule. Early stopping and learning rate reduction strategies are employed to mitigate overfitting. The best model weights for each fold are saved based on the fused-stream validation accuracy.

The loss function adopts a multi-headed structure that incorporates the contributions of the fused, visual, and auditory outputs. The auxiliary visual and auditory outputs are included only during training with small weights to enhance the stability of the learned representations. This formulation enables more balanced learning centered around the primary fusion output.

In summary, the GCM-AFN architecture is capable of effectively learning from both visual and auditory information and integrates complementary cross-stream cues through its gated fusion mechanism. Consequently, the model achieves high performance in terms of both accuracy and generalization in the scene classification task.

3. RESULTS

The performance of the proposed Gated Cross-Modal Attention Fusion Network (GCM-AFN) architecture was evaluated using a 10-fold cross-validation protocol. For each fold, the training process was conducted for 100 epochs, and the best model weights were selected based on the fused-stream accuracy obtained on the validation set.

The averaged loss and accuracy curves from the training and validation processes across the 10 folds are presented in Figure 3. The figure illustrates the mean trends for each epoch and the variability across folds, represented by the $\pm$ standard deviation bands.

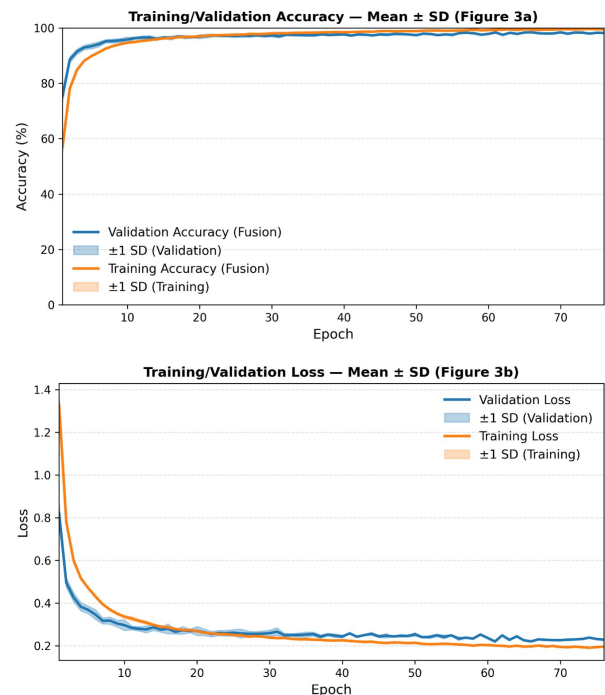


Figure 3. Mean training and validation accuracy (a) and loss (b) curves of the GCM-AFN model across 10-fold cross-validation. The curves show the average values across all folds along with the $\pm$ standard deviation intervals.

The overall pattern of the training and validation loss curves indicates that the model begins to gain rapid learning momentum after approximately the fifth epoch. This period corresponds to the stage at which the pretrained EfficientNetV2-B1 backbone starts to be gradually unfrozen. Between roughly epochs 17 and 23, the validation performance reaches a plateau, after which the fluctuations diminish and the model transitions into a stable learning phase. The gradual reduction of the learning rate through cosine decay and the ReduceLROnPlateau strategy enhances model stability and limits the tendency toward overfitting.

The accuracy values obtained across the 10-fold cross-validation procedure are presented in Table 2. The table reports the accuracy rates for the fused (combined), auditory, and visual streams for each fold.

Table 2. Fusion, auditory, and visual accuracy values for each fold of the GCM-AFN model in the 10-fold cross-validation

Fold	Fusion Accuracy (%)	Audio Stream (%)	Visual Stream (%)
1	98.15	97.6	97.1
2	98.15	97.4	97.0
3	98.26	97.7	97.3
4	98.03	97.2	96.8
5	98.43	97.9	97.3
6	97.22	96.6	96.1
7	97.62	96.9	96.6
8	98.20	97.6	97.1
9	97.86	97.2	96.7
10	98.03	97.4	96.9
Mean $\pm$ SS	97.99 $\pm$ 0.37	97.5 $\pm$ 0.4	97.2 $\pm$ 0.5

As shown in the table, the model achieves high and consistent accuracy across all folds. The fused configuration attains an average accuracy of $97.99 \pm 0.37\%$, outperforming both the auditory stream ($97.5 \pm 0.4\%$) and the visual stream ($97.2 \pm 0.5\%$). The combined use of visual and auditory information provides an average improvement of approximately 1.8% over the single-stream models, demonstrating that the architecture effectively leverages complementary information from both sources.

To evaluate the class-wise performance of the model, the class reports obtained from each fold were averaged. These results are summarized in Table 3.

Table 3. Class-wise average precision, recall, and F1-score values of the GCM-AFN model

Class	Precision	Recall	F1-score
Classroom	0.990 ± 0.003	0.989 ± 0.004	0.989 ± 0.003
City	0.973 ± 0.007	0.959 ± 0.008	0.966 ± 0.007
River	0.949 ± 0.010	0.955 ± 0.008	0.952 ± 0.009
Beach	0.964 ± 0.006	0.969 ± 0.006	0.966 ± 0.006
Grocery-store	0.979 ± 0.004	0.975 ± 0.005	0.977 ± 0.004
Football-match	0.982 ± 0.004	0.983 ± 0.004	0.982 ± 0.004
Restaurant	0.937 ± 0.012	0.931 ± 0.015	0.934 ± 0.013
Jungle	0.994 ± 0.002	0.995 ± 0.002	0.994 ± 0.002
Forest	1.000 ± 0.000	0.998 ± 0.001	0.999 ± 0.001
Macro Average	0.974	0.972	0.973

As shown in the table, the model achieves a strong balance across classes. The macro F1-score is approximately 0.973. The highest performance is observed in the Forest, Jungle, and Classroom classes ($F1 > 0.99$), while the lowest performance appears in the Restaurant class ($F1 \approx 0.93$). This indicates that the model experiences slightly more difficulty in scenes with higher environmental audio-visual complexity, yet it still demonstrates strong overall generalization capability.

To examine inter-class confusion patterns, a normalized confusion matrix was constructed by averaging the results obtained from all folds. Figure 4 provides a visual

representation of the correct and misclassified proportions for each class.

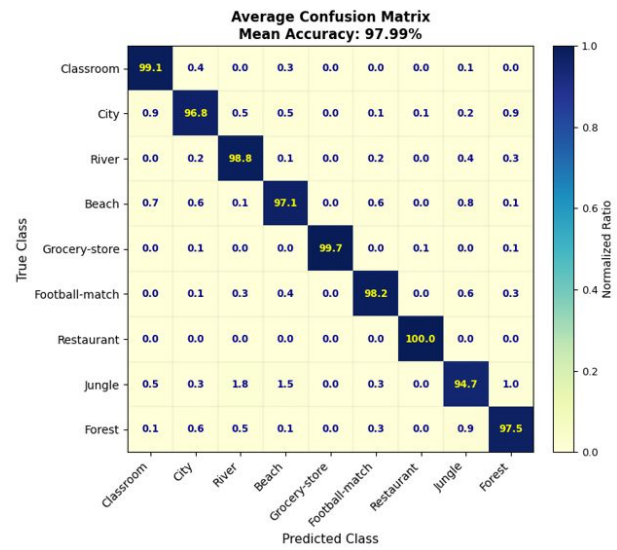


Figure 4. Average normalized confusion matrix of the GCM-AFN model. Each cell represents the prediction rate for the corresponding class.

An examination of the confusion matrix reveals that misclassifications predominantly occur between scenes that are visually or acoustically similar. In particular, Restaurant samples are occasionally confused with Classroom, City, and Football-match scenes, while limited cross-class predictions are observed among the River–Beach, Forest–Jungle, and City–River pairs. Overall, the model achieves approximately 1.5–2.0% higher accuracy compared to the individual visual or auditory streams, demonstrating consistent generalization performance across all subsets of the dataset.

4. DISCUSSION

In this study, the proposed method was evaluated on the extended Kaggle version of the Scene Classification – Images and Audio dataset, which contains 17,252 samples across nine classes. Earlier versions of the same dataset were examined using different fusion strategies by Bird et al. [11] and Woodward and Kanjo [22].

In the study by Bird et al. [11], a pretrained VGG16 network was used for the visual stream, while a multilayer perceptron (MLP) model based on 13 MFCC features was applied to the auditory stream. The two streams were combined using a late-fusion fully connected layer. Their 10-fold cross-validation results reported an average accuracy of 96.81%.

Woodward and Kanjo [22] proposed a lightweight early-feature fusion model designed for embedded systems using the Kaggle version of the dataset. The visual stream was processed with a 2D CNN, the auditory stream with a 1D CNN, and the extracted features were compressed in parallel and fed into a small MLP. Using an 80/20 train–test split, the authors reported a fusion accuracy of 93.0%.

The proposed Gated Cross-Modal Attention Fusion Network (GCM-AFN) architecture learns the interaction between audio and visual streams through attention-guided feature integration. Under the 10-fold cross-validation protocol, the model achieves an average accuracy of $97.99 \pm 0.37\%$, and the low variance indicates stable generalization across different subsets of the data.

In Bird et al. [11], the original eight-class version of the dataset was used. Although Woodward and Kanjo [22] stated that the dataset contained 16,000 samples, their class list includes nine environmental categories, suggesting that the Kaggle version may have been used implicitly. In the present study, the extended nine-class Kaggle version is explicitly evaluated. A performance comparison between previous studies and the proposed approach is presented in Table 4.

Table 4. Accuracy comparison of studies conducted on the Look and Listen dataset.

Study	Dataset(Number of Samples)	Evaluation Protocol	Accuracy (%)
Bird et al. [11]	16000-8 class	10-Fold CV	96.81
Woodward & Kanjo [22]	16000- 9 class	80/20 hold-out	93.00
This Study (GCM-AFN)	17252-9 class	10-Fold CV	97.99±0.37

The obtained results demonstrate that integrating the features extracted from the audio and visual streams through an attention mechanism substantially enhances both scene classification performance and model stability. The GCM-AFN architecture achieves high accuracy with low variance across different subsets of the data, effectively leveraging information from both sources. Although direct comparison with earlier studies using the original eight-class dataset is somewhat limited, the findings indicate that attention-based integration provides a more stable and generalizable information transfer mechanism for multimodal scene recognition tasks.

In this study, a 10-fold stratified cross-validation strategy was employed to ensure reliable performance evaluation. However, evaluating the generalization capability of the proposed model across different datasets and data sources remains a limitation and should be addressed in future studies.

5. CONCLUSION

In this study, a new multi-stream deep learning architecture, the Gated Cross-Modal Attention Fusion Network (GCM-AFN), was presented for jointly analyzing visual and auditory information in scene classification tasks. By integrating deep representations extracted from each stream through a gated attention mechanism, the proposed model dynamically learns cross-stream interactions and produces more coherent representations of complementary scene cues.

The 10-fold cross-validation experiments demonstrate that GCM-AFN surpasses both single-stream models and existing early- and late-fusion approaches, achieving higher accuracy and lower variance. The approximately 98% accuracy and strong class-wise scores indicate that

the model possesses robust generalization capability across different scene types. Confusion matrix analyses further confirm that the model can reliably distinguish scenes that exhibit high visual or auditory similarity.

These findings validate that attention-guided fusion provides an effective and stable information integration strategy for visual-auditory scene classification. Future work will focus on evaluating the approach under varying environmental conditions, incorporating more diverse audio-visual datasets, and exploring its applicability within broader real-world scenarios.

REFERENCES

- [1] Çelik Y. Application of deep learning for voice command classification in Turkish language. Bitlis Eren University Journal of Science. 2024;13(3):701–708. doi:10.17798/bitlisfen.1477191.
- [2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM. 2017;60(6):84–90. doi: 10.1145/3065386.
- [3] Güneş H, Hark C, Akkaya AE. Comparison of Deep Learning Models and Optimization Algorithms in the Detection of Scoliosis and Spondylolisthesis from X-Ray Images. Sakarya University Journal of Science. 2024;28(2):438–451. doi:10.16984/saufenbilder.1246001.
- [4] Doğan F, Aktaş M, Gürsoy Mİ. Classification of Skin Diseases with Different Deep Learning Models and Comparison of the Performances of the Models. TDFD. 2024;13(3):117–123. doi:10.46810/tdfd.1502471.
- [5] Ceylan T, İnik Ö. Development of an Effective Deep Learning Model for COVID-19 Detection from CT Images. Tr. J. Nature Sci. 2025;14(1):156–166. doi:10.46810/tdfd.1472034.
- [6] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv Preprint. 2013;arXiv:1301.3781.
- [7] Liu X, Wang H, Li Z, Qin L. Deep learning in ECG diagnosis: A review. Knowledge-Based Systems. 2021;227:107187. doi: 10.1016/j.knosys.2021.107187.
- [8] İnik Ö. Classification of Scenes in Aerial Images with Deep Learning Models. TDFD. 2023;12(1):37–43. doi:10.46810/tdfd.1225756.
- [9] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444. doi: 10.1038/nature14539.
- [10] Uçkan T, Aslan C, Hark C. A Comprehensive Hybrid Approach for Indoor Scene Recognition Combining CNNs and Text-Based Features. Sensors. 2025;25(17):5350. doi: 10.3390/s25175350.
- [11] Bird JJ, Ekárt A, Faria DR. Look and listen: A multi-modality late fusion approach to scene classification for autonomous machines. In: Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems

- (IROS). IEEE; 2020. p. 10555–10560. doi: 10.1109/IROS45743.2020.9341557.
- [12] Wang S, Heittola T, Mesáros A, Virtanen T. Audio-visual scene classification: Analysis of DCASE 2021 challenge submissions. arXiv Preprint. 2021;arXiv:2105.13675. doi: 10.48550/arXiv.2105.13675.
- [13] Lei H, Chen N. Audio-visual scene classification based on multi-modal graph fusion. In: Proceedings of INTERSPEECH 2022. 2022. p. 4157–4161. doi: 10.21437/interpeech.2022-741.
- [14] Chen C, Wang M, Zhang P. Audio-visual scene classification using a transfer-learning-based joint optimization strategy. arXiv Preprint. 2022;arXiv:2204.11420. doi: 10.48550/arXiv.2204.11420.
- [15] Pham L, Ngo D, Nguyen T, Nguyen P, Hoang T, Schindler A. An audio-visual dataset and deep learning frameworks for crowded scene classification. In: Proceedings of the 19th International Conference on Content-Based Multimedia Indexing (CBMI). ACM; 2022. p. 23–28. doi: 10.1145/3549555.3549568.
- [16] Zhou L, Zhou Y, Qi X, Hu J, Lam TL, Xu Y. Feature pyramid attention network for audio-visual scene classification. CAAI Transactions on Intelligence Technology. 2025;10(2):359–374. doi: 10.1049/cit2.12375.
- [17] Xu C, Xu L, Zhao S, Yu L, Zhang C. Complementary knowledge-augmented multimodal learning method for yarn quality soft sensing. Engineering Applications of Artificial Intelligence. 2024;133:108057. doi: 10.1016/j.engappai.2024.108057.
- [18] Liang X, Qian Y, Guo Q, Cheng H, Liang J. AF: An association-based fusion method for multimodal classification. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021;44(12):9236–9254. doi:10.1109/TPAMI.2021.3125995.
- [19] Cantarini M, Gabrielli L, Mancini A, Squartini S, Longo R. A3CarScene: An audio-visual dataset for driving scene understanding. Data in Brief. 2023;48:109146. doi:10.1016/j.dib.2023.109146.
- [20] Sumathi K, Kumar P, Mahadevaswamy HR, Ujwala BS. Optimizing multimodal scene recognition through relevant feature selection approach for scene classification. MethodsX. 2025;14:103226. doi:10.1016/j.mex.2025.103226.
- [21] Bird JJ. Scene Classification – Images and Audio [Data set]. Kaggle. Available from: <https://www.kaggle.com/datasets/birdy654/scene-classification-images-and-audio>
- [22] Woodward K, Kanjo E. Parallel Feature Fusion for Multimodal Scene Recognition on Dual-Core MCUs. IEEE Pervasive Comput. 2025;24(3):54–62. doi: 10.1109/mprv.2025.3555334.