applied sciences

MDPI

Article

Optimizing Multimodal Scene Recognition through Mutual Information-Based Feature Selection in Deep Learning Models

Mohamed Hammad [1,2,*], Samia Allaoua Chelloug [3,*], Walaa Alayed [3] and Ahmed A. Abd El-Latif [1,4,5]

[1] EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia; aabdellatif@psu.edu.sa
[2] Department of Information Technology, Faculty of Computers and Information, Menoufia University, Shebin El Kom 32511, Egypt
[3] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia; wmalayed@pnu.edu.sa
[4] Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), Besut Campus, Besut 22200, Malaysia
[5] Department of Mathematics and Computer Science, Faculty of Science, Menoufia University, Shebin Elkom 32511, Egypt
* Correspondence: mhammad@psu.edu.sa (M.H.); sachelloug@pnu.edu.sa (S.A.C.)

Abstract: The field of scene recognition, which lies at the crossroads of computer vision and artificial intelligence, has experienced notable progress because of scholarly pursuits. This article introduces a novel methodology for scene recognition by combining convolutional neural networks (CNNs) with feature selection techniques based on mutual information (*MI*). The main goal of our study is to address the limitations inherent in conventional unimodal methods, with the aim of improving the precision and dependability of scene classification. The focus of our research is around the formulation of a comprehensive approach for scene detection, utilizing multimodal deep learning methodologies implemented on a solitary input image. Our work distinguishes itself by the innovative amalgamation of CNN- and *MI*-based feature selection. This integration provides distinct advantages and enhanced capabilities when compared to prevailing methodologies. In order to assess the effectiveness of our methodology, we performed tests on two openly accessible datasets, namely, the scene categorization dataset and the AID dataset. The results of these studies exhibited notable levels of precision, with accuracies of 100% and 98.83% achieved for the corresponding datasets. These findings surpass the performance of other established techniques. The primary objective of our end-to-end approach is to reduce complexity and resource requirements, hence creating a robust framework for the task of scene categorization. This work significantly advances the practical application of computer vision in various real-world scenarios, leading to a large improvement in the accuracy of scene recognition and interpretation.

Keywords: scene recognition; convolutional neural network (CNN); deep learning; mutual information (*MI*); multimodel; AID

Citation: Hammad, M.; Chelloug, S.A.; Alayed, W.; El-Latif, A.A.A. Optimizing Multimodal Scene Recognition through Mutual Information-Based Feature Selection in Deep Learning Models. *Appl. Sci.* **2023**, *13*, 11829. https://doi.org/10.3390/app132111829

Academic Editor: Fan Zhang

Received: 7 October 2023
Revised: 26 October 2023
Accepted: 27 October 2023
Published: 29 October 2023

1. Introduction

Scene recognition, which is a crucial aspect of computer vision and artificial intelligence, has been the subject of extensive research across multiple disciplines and has received significant interest in recent decades [1]. This discipline centers on the advancement of algorithms, models, and procedures that possess the ability to classify and comprehend the content of photographs or videos by analyzing the scenes they portray automatically and precisely [2]. Recognition of scenes is of great significance in various domains, encompassing autonomous navigation, retrieval of images and videos, robotics, surveillance, and augmented reality, among other fields [3–7]. The main goal of scene recognition is to provide machines with the capacity to understand the visual world in a manner similar to

humans through the recognition and interpretation of intricate environmental scenarios [8]. This involves the identification of different components present in a given scene, including objects, spatial arrangements, and contextual associations, followed by the deduction of the scene's overarching semantic significance. The attainment of this particular talent is a multifaceted undertaking that involves various obstacles, such as comprehending scenes, detecting and recognizing objects, modeling context, and segmenting images.

Traditionally, the field of scene recognition has predominantly employed a unimodal methodology, wherein image or video input is analyzed using a solitary model [9]. Recently, there has been a notable increase in research attention and progress in the field of scene identification [10–20]. This rise can be attributed to the advancements made in deep learning techniques, the widespread availability of extensive datasets with annotations, and the accessibility of high-performance computer resources. CNNs and other deep learning architectures have exhibited notable achievements in the domain of scene recognition tasks, outperforming conventional computer vision techniques in both accuracy and efficiency. The aforementioned technological advancements have facilitated the development of novel applications, such as autonomous vehicles, intelligent surveillance systems, and augmented reality encounters, which heavily depend on the comprehension and analysis of visual scenes. In recent times, multimodal deep learning techniques have demonstrated remarkable success in scene recognition tasks, surpassing the capabilities of traditional unimodal approaches [21–23]. These advances have paved the way for innovative applications such as self-driving cars, smart surveillance systems, and augmented reality experiences, all of which rely on scene understanding and interpretation. These methodologies make use of several sources of information, including text, audio, and depth data, in order to improve the comprehension of scenes. This paper centers on the utilization of multimodal deep learning techniques for scene recognition. The objective is to leverage the advantages of different modalities in order to enhance the robustness and accuracy of the findings obtained.

The core of our methodology revolves around the novel idea of integrating many modalities into a single input for scene recognition. This approach deviates from traditional models that predominantly rely on processing data from a single modality. This innovative methodology utilizes multimodal deep learning on the identical input image data, facilitating the concurrent analysis of many forms of information, such as text, audio, and depth data, all derived from the same image. In this particular framework, we effectively incorporate many modalities into a unified model, leveraging their complementary nature to attain a more thorough comprehension of the scenario. In order to enhance the quality of the feature representations resulting from this fusion, we utilize the technique of *MI*-based feature selection [24]. The utilization of this particular technique results in an improvement in the overall quality, discriminative capability, and resilience of the features, thereby providing substantial advantages for tasks related to scene identification. The last stage of recognition is performed by the SoftMax layer within our integrated model. This layer utilizes the information acquired from our innovative single-input multimodal technique to assign scene labels. The methodology employed in our study capitalizes on the advantageous effects of multimodal fusion when applied to identical input image data, resulting in significantly enhanced accuracy and resilience in scene recognition. The novel contributions of our research are outlined below:

- Our approach differs from previous models that primarily focused on single-modal processing by introducing a multimodal fusion technique that works on the same input visual data. The present methodology introduces a novel approach that allows for the simultaneous processing of many types of information derived from a singular image. This advancement facilitates a more thorough comprehension of the scene compared to prior research that predominantly concentrated on single-modal techniques.
- To enhance feature quality and robustness, we employ *MI*-based feature selection. Our approach incorporates *MI* not only for feature selection but also within the context of multimodal fusion. We utilize *MI* to assess the interdependence between different

data modalities and scene labels, enabling effective fusion strategies that are tailored to each modality's relationship with scene categories.

- The model incorporates an additional layer for integrating optimized features extracted from the input image data, representing an innovative architectural element. This feature integration step significantly enhances the quality of feature representations, further contributing to the model's superior performance compared to techniques that do not incorporate such integration.
- Our method introduces the novel concept of an end-to-end model, eliminating the need for complex multistage pipelines typically employed in traditional scene recognition approaches.

Our multimodal fusion scene recognition method is presented in various key sections of the paper. After the Introduction, the literature review is discussed in Section 2. The methodology describes our model's essential components in Section 3. We then give empirical results and explain their consequences in Section 4. Our succinct conclusion highlights our approach's contributions and transformative possibilities in scene recognition in Section 5.

2. Literature Review

Extensive investigation has been carried out in the field of scene recognition, encompassing several procedures and techniques with the objective of progressing the current level of knowledge. This section examines the progression of scene recognition methodologies, the advent of deep learning architectures, and the advancements made by multimodal fusion approaches. This section also offers significant contextual information for comprehending the existing body of research, emphasizing notable discoveries, obstacles, and the progression of improvements in the subject.

Hua et al. [10] presented a prototype memory network to detect many scenes in one image. This was conducted by using several well-annotated images of individual scenarios. The network has three main components: a prototype learning module, an external memory that can store prototypes, and a multihead attention memory retrieval module. They obtained the best mean *F1* score of 57.40% using the AID dataset. Petrovska et al. [11] suggested fine-tuning pre-trained CNNs for end-to-end aerial scene classification using transfer learning. Feature extraction from fine-tuned neural networks is followed by remote sensing picture categorization with a Support Vector Machine (SVM) model with linear and radial basis function (RBF) kernels. To reduce overfitting in pre-trained models, label smoothing regularization is used. Inception-v3, Xception, Residual Network50 (ResNet50), and DenseNet121, inception- and residual-based CNNs, are used for fine-tuning and feature extraction. The method has a classification accuracy of up to 98%, outperforming other methods. Wang et al. [12] developed a depth-wise separable convolution (CSDS) network-based channel–spatial attention mechanism for aerial scene categorization. The researchers first created DS-Conv and pyramid residual connection architectures. The DS-Conv approach efficiently recovers and combines channel features, reducing processing needs. Pyramid residual connections also link features from multiple layers, enabling relationships. They use an upgraded cross-entropy loss function to reduce the influence of comparable categories during backpropagation. They obtained an accuracy of 94.29% using the AID dataset. Zhao et al. [13] used channel–spatial attention in a residual dense network architecture to classify remote sensing images. Initially, leftover dense blocks are used to fuse multilayer convolutional features. A channel–spatial attention module is included for better feature representation. After data augmentation, the SoftMax classifier classifies the scene. They obtained an overall accuracy of 94.15% using the AID dataset. Bazi et al. [14] presented a method for remote sensing image classification using vision transforms. They first split the images into smaller portions, which are flattened and embedded sequentially. The resulting sequence is then sent into numerous multiheaded attention layers to create the final representation. The initial token sequence is fed into a SoftMax classification layer. The researchers study numerous data augmentation methods

to generate more training data to improve classification performance. Researchers classified the AID dataset with 95.86% accuracy. Wang and Yu [15] presented the extraction of two informative feature categories from the raw RGB and saliency-coded network streams. The authors propose a deep feature fusion model that combines these two sets of features for classification. They obtained an accuracy of 93.70% using the AID dataset. Wu et al. [16] presented S-MAT, a semantic-driven Masked Attention Transformer method for multilabel aerial scene image categorization. S-MAT uses a Masked Attention Transformer (MAT) to capture correlations in Semantic Disentanglement Module-generated label embeddings. In the presented MAT, the masked attention method eliminates needless dependencies and improves model resilience. Thus, the strategy presented in this work may directly and properly capture label interdependence. Thus, the researchers' method yields classification *F1* scores of 90.90% on the AID dataset.

These previous methods have made significant contributions to scene recognition. However, some of these methods may struggle with more complex scenes due to their reliance on individual prototypes, such as [10,12,15,16]. Also, authors in [10–13] used fine-tuning approaches, which can be prone to overfitting, and the use of SVM models for classification may not capture complex interdependencies. Several methods might still face challenges in handling scene variations, such as [14,15]. For other methods, such as [14,16], the performance might be affected by the effectiveness of data augmentation methods. Our work seeks to address these limitations by introducing a novel single-input multimodal fusion approach. This method leverages multiple modalities from the same image, enabling it to handle complex scenes effectively. Moreover, we employ *MI*-based feature selection to enhance feature quality and discriminative capabilities, further improving scene recognition accuracy and robustness. Our approach offers a comprehensive solution that aims to overcome the limitations encountered in prior studies, making it a promising avenue for advancing scene recognition.

3. Methodology and Materials

Our methodology centers on the innovative concept of single-input multimodal fusion for scene recognition, departing from traditional single-modal approaches. Figure 1 shows the block diagram of the proposed multi-deep model for scene recognition. As shown in the figure, the input images are fed directly to the proposed multi-deep model. Then, the features are extracted from the first fully connected layer and fused using an addition technique. To optimize feature representations, we employ MI-based feature selection. These refined features are then incorporated as an additional layer into the model, and recognition is performed using the SoftMax layer. An additional layer in the model greatly improves scene identification feature relevancy. Through a systematic evaluation of feature significance and selection of the most relevant features, we can improve model accuracy and efficiency by lowering feature space dimensionality. This approach harnesses synergistic information from multiple modalities, improving feature quality and scene recognition performance. All steps of the model are discussed in the following subsections.

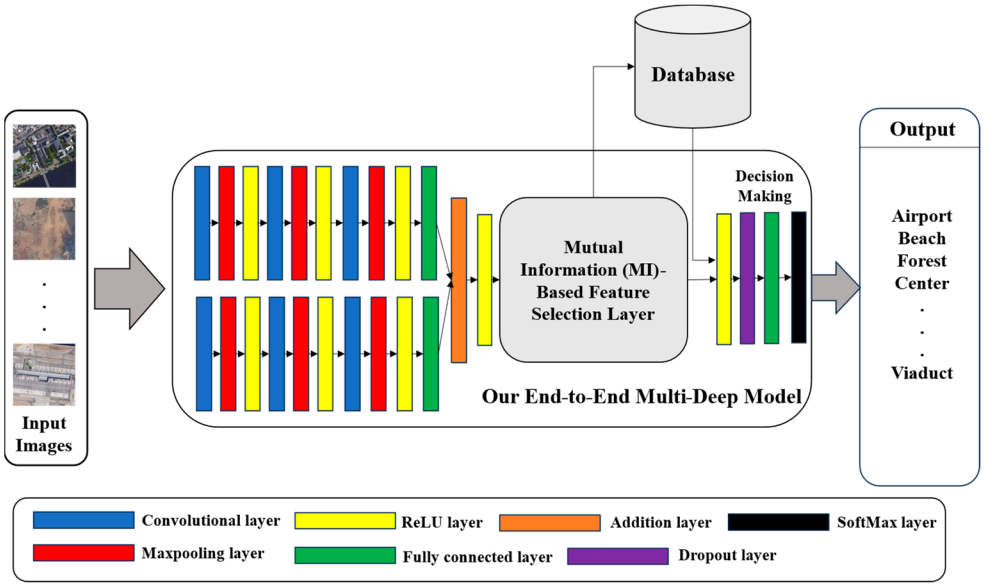

Figure 1. General block diagram of our multi-deep model.

3.1. Our Multimodal Deep Learning Structure

The presented deep learning structure for scene recognition showcases a comprehensive architecture designed to extract and harness informative features from multimodal data sources. The architecture is characterized by a series of convolutional and pooling layers, which are crucial for feature extraction and dimensionality reduction, as shown in Figure 2. The convolutional layers perform spatial filtering to capture relevant patterns in the input data. Following each convolutional layer, max-pooling layers downsample the feature maps, enhancing computational efficiency while preserving essential information. Rectified Linear Unit (ReLU) layers introduce non-linearity, enabling the model to capture complex relationships within the data. A summary of the layers and filter sizes of our multimodal deep learning model is shown in Table 1.

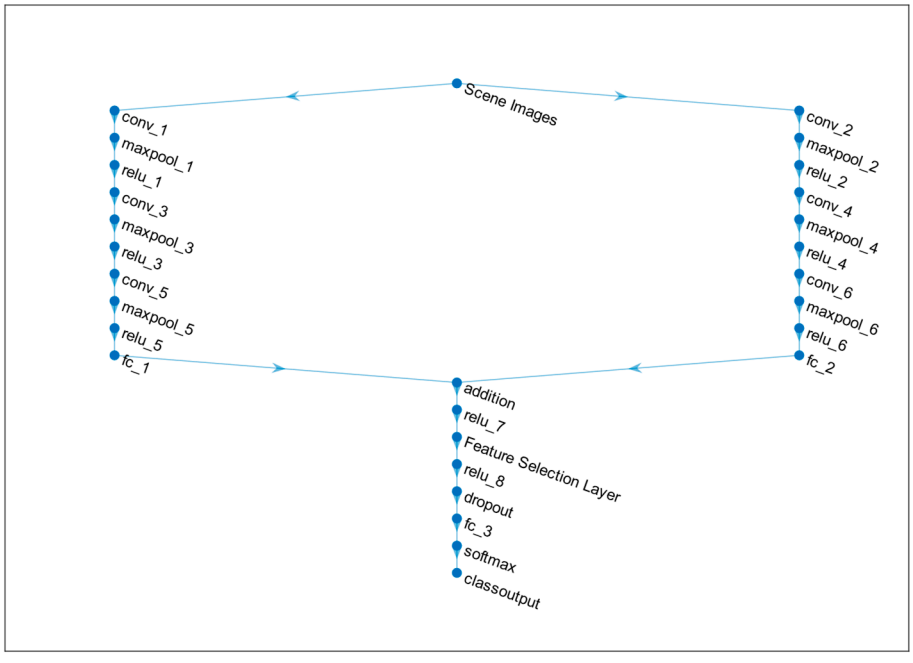

Figure 2. Structure of the proposed multi-deep model.

Table 1. Summary of our multimodal deep learning model.

	Name	Type	Activation	Learnables
1	Scene Images 200 × 200 × 3	Image Input	200 × 200 × 3	-
2	Conv_1 1283 × 3 × 3 with stride [1 1] and padding 'same'	Convolution	200 × 200 × 128	Weights 3 × 3 × 3 × 128 Bias 1 × 1× 128
3	Maxpool_1 2 × 2 with padding 'same'	Max Pooling	200 × 200 × 128	-
4	Relu_1 ReLU	ReLU	200 × 200 × 128	-
5	Conv_3 643 × 3 × 128 with stride [1 1] and padding 'same'	Convolution	200 × 200 × 64	Weights 3 × 3 × 128 × 64 Bias 1 × 1× 64
6	Maxpool_3 2 × 2 with padding 'same'	Max Pooling	200 × 200 × 64	-
7	Relu_3 ReLU	ReLU	200 × 200 × 64	-
8	Conv_5 323 × 3 × 64 with stride [1 1] and padding 'same'	Convolution	200 × 200 × 32	Weights 3 × 3 × 64 × 32 Bias 1 × 1 × 32
9	Maxpool_5 2 × 2 with padding 'same'	Max Pooling	200 × 200 × 32	-
10	Relu_5 ReLU	ReLU	200 × 200 × 32	-
11	Fc_1 1024	Fully Connected	1 × 1 × 1024	Weights 1024 × 1,280,000 Bias 1024 × 1
12	Conv_2 1283 × 3×3 with stride [1 1] and padding 'same'	Convolution	200 × 200 × 128	Weights 3 × 3 × 3 × 128 Bias 1 × 1 × 128
13	Maxpool_2 2 × 2 with padding 'same'	Max Pooling	200 × 200 × 128	-
14	Relu_2 ReLU	ReLU	200 × 200 × 128	-
15	Conv_4 643 × 3 × 128 with stride [1 1] and padding 'same'	Convolution	200 × 200 × 64	Weights 3 × 3 × 128 × 64 Bias 1 × 1 × 64
16	Maxpool_4 2 × 2 with padding 'same'	Max Pooling	200 × 200 × 64	-
17	Relu_4 ReLU	ReLU	200 × 200 × 64	-
18	Conv_6 323 × 3 × 64 with stride [1 1] and padding 'same'	Convolution	200 × 200 × 32	Weights 3 × 3 × 64 × 32 Bias 1 × 1 × 32
19	Maxpool_6 2 × 2 with padding 'same'	Max Pooling	200 × 200 × 32	-
20	Relu_6 ReLU	ReLU	200 × 200 × 32	-
21	Fc_2 1024	Fully Connected	1 × 1 × 1024	Weights 1024 × 1,280,000 Bias 1024 × 1
22	Addition Eliminate-wise addition of 2 input	Addition	1 × 1 × 1024	-

Table 1. *Cont.*

	Name	Type	Activation	Learnables
23	Relu_7 ReLU	ReLU	1 × 1 × 1024	-
24	Feature Selection Layer 600 fully connected	Fully Connected	1 × 1 × 600	Weights 600 × 1024 Bias 600 × 1
25	Relu_8 ReLU	ReLU	1 × 1 × 600	-
26	Dropout 50% dropout	Dropout	1 × 1 × 600	-
27	Fc_3 6	Fully Connected	1 × 1 × 6	Weights 6 × 600 Bias 6 × 1
28	Softmax	Softmax	1 × 1 × 6	-
29	Classoutput Crossentropyex with 'Buildings' and other 5 classes	Classification output	-	-

One of the notable features of this architecture is the incorporation of two parallel pathways for feature extraction. These pathways allow for the simultaneous processing of multimodal data. By integrating features from diverse modalities at an early stage, the model can effectively capture complementary information, enhancing its scene recognition capabilities. The architecture employs an 'additionLayer' ('addition') to merge features from both pathways, followed by an additional ReLU layer. This fusion step is pivotal for effectively combining the extracted features from different modalities. Subsequently, a 'Feature Selection Layer' ('Feature Selection Layer') introduces a critical element of feature selection. This layer, which follows the ReLU activation, uses a fully connected layer with 600 units to select the most relevant features from the fused representation. Feature selection is a crucial mechanism for enhancing the model's interpretability and reducing computational complexity. After feature selection, a dropout layer ('dropout') helps prevent overfitting by randomly deactivating a portion of the selected features. Following this, another fully connected layer is responsible for mapping the feature representation to the output space, and a SoftMax layer ('softmax') produces probability distributions over scene categories. Finally, the 'classificationLayer' ('classoutput') assigns scene labels based on the probability distribution, facilitating scene recognition.

The architecture's training is guided by specific options, including the choice of the Adam optimizer, mini-batch size, maximum epochs, and learning rate scheduling. Training progress is visualized using training plots, and the model's performance is validated using a separate validation dataset. All hyperparameters are listed in Table 2.

Table 2. Hyperparameters of our deep multimodal model.

Hyperparameter	Value
Optimizer	Adam
Mini-Batch Size	4
Kernel Size	3 × 3 for Convolution 2 × 2 for Max pooling
Number of kernels	Conv_1 and Conv_2 = 128 Conv_3 and Conv_4 = 64 Conv_5 and Conv_6 = 32
Number of Nodes in FC layers	Fc_1 and Fc_2 = 1024 Fc_3 = 6

Table 2. *Cont.*

Hyperparameter	Value
Maximum Epochs	30
Initial Learning Rate	3.0000000×10^{-4}
Learning Rate Schedule	Piecewise
Learning Rate Drop Factor	0.5
Learning Rate Drop Period	5
Data Shuffling	Every Epoch
Validation Frequency	87

3.2. MI for Feature Selection

MI is a fundamental concept in information theory that plays a pivotal role in various fields, including scene recognition in computer vision. *MI* quantifies the dependency between two random variables, providing a measure of the amount of information that knowing one variable provides about the other. In the context of scene recognition, *MI* is a valuable tool for assessing the relationship between different image features or modalities, aiding in feature selection, fusion, and optimization processes.

Mathematically, *MI* between two discrete random variables X and Y is defined as [25]:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

where $MI(X;Y)$ represents the *MI* between X and Y. $p(x,y)$ is the joint probability mass function of X and Y. $p(x)$ and $p(y)$ are the marginal probability mass functions of X and Y, respectively. In essence, *MI* quantifies how much knowing one variable reduces the uncertainty about the other variable. If X and Y are independent, $MI(X;Y)$ equals zero, as the knowledge of one variable provides no information about the other. Conversely, when X and Y are perfectly dependent, $MI(X;Y)$ is maximized.

In the context of scene recognition, *MI* is utilized to assess the degree of dependency or correlation between image features or modalities. This evaluation is pivotal in several aspects of scene recognition:

(1) Feature Selection: *MI* helps identify and select the most informative features or modalities for scene recognition. By calculating *MI* between each feature (F) and the target scene label (L), less relevant or redundant features can be pruned, enhancing computational efficiency and reducing the risk of overfitting. The formula for this feature selection process is:

$$MI(F,L) = \sum (f \in F) \sum (l \in L) p(f,l) * log\left[\frac{p(f,l)}{p(f) * p(l)}\right] \quad (2)$$

where

- $MI(F,L)$ represents the *MI* between the feature F and the scene label L.
- $p(f,l)$ is the joint probability mass function of the feature F and the scene label L.
- $p(f)$ and $p(l)$ are the marginal probability mass functions of the feature F and the scene label L, respectively.

(2) Multimodal Fusion: In multimodal scene recognition, *MI* quantifies the interdependence between different data modalities (e.g., text, audio, depth) and the scene labels (L). This assists in determining how to combine modalities effectively, such as weighing them based on their *MI* with the scene label. The weighted fusion formula can be represented as:

$$Weighted\ Fusion = W1 * MI(M1,L) + W2 * MI(M2,L) + \ldots \quad (3)$$

where

- Weighted Fusion represents the combined information from multiple modalities.
- $MI(Mi,L)$ is the mutual information between each modality (Mi) and the scene label L.
- $w1$, $w2$, etc., are weights assigned to each modality, which can be determined based on their MI values.

(3) Optimization: MI can be employed in optimization processes to improve feature representations. Genetic Algorithms, for example, can utilize MI as an objective function, enhancing the discriminative power and robustness of features extracted from the same input image data. The objective function for optimization can be defined as:

$$Objective\ Function = MI(F1, L) + MI(F2, L) + \ldots + MI(Fn, L) \quad (4)$$

where

- $MI(Fi,L)$ represents the mutual information between each feature (Fi) and the scene label L.
- n is the number of features under consideration.

In our study, MI served as a crucial tool for feature selection in the context of scene recognition. Feature selection is a pivotal step in enhancing the efficiency and accuracy of scene recognition models. The primary aim of MI-based feature selection is to systematically evaluate the relevance of individual features extracted from the proposed deep multimodal architecture. These features originate from the first fully connected layer of both models, each representing distinctive characteristics of the input images. MI, in this paper, quantifies the degree of information shared between each feature and its correlation with scene labels. The investigation of several techniques for selecting features is a significant area of scholarly inquiry [26,27]. For instance, Liu et al. [26] tested three upgraded feature selection algorithms on numerous Chinese text categories. This article discusses CHMI, an improved version of the CHI square with MI method that includes word frequency and term correction. TF–CHI adds phrase frequency to the weight calculation. Finally, the extreme gradient boosting (XGBoost) technique is added to the TF–IDF algorithm to improve word filtering. The empirical results in their work show that the feature selection algorithms improve performance across varied news corpora. However, each corpus type has different optimal feature selection strategies. In our work, the selection of MI as the chosen method is justified based on its strong theoretical foundation, demonstrated empirical efficacy, and its suitability for the specific demands of scene recognition. It is crucial to acknowledge that the selection of MI was not made in isolation but rather through a thorough evaluation of its advantages and its alignment with the research aims.

The input for MI-based feature selection in our model consists of the extracted feature vector of the two models within the multimodal architecture. These features, derived from distinct modalities, collectively capture rich information about the input images. Additionally, the input includes scene labels that categorize these images into various scene classes. The output of the MI-based feature selection process is a refined feature set. This subset is composed of features with high MI scores, indicating a substantial mutual dependency between these selected features and the scene labels. Notably, the MI-based feature selection includes an added layer after the fusion of the two models, responsible for selecting suitable features from both models and combining them into one vector. This combined feature vector is then integrated into the deep model to complete the feature selection process. To implement MI-based feature selection in our model, we adopt the following algorithmic steps:

- As an initial step, we gather a dataset comprising feature vectors from the first fully connected layer of both models in our multimodal architecture. This dataset also includes the corresponding scene labels.

- For each feature in the dataset, *MI* is calculated with respect to the scene labels, following the *MI* formula. Importantly, this *MI* calculation takes place after the fusion of the two models.
- Based on the computed *MI* scores, we rank the features in descending order. Higher *MI* scores signify stronger *MI* content between selected features and scene labels.
- The top N features with the highest *MI* scores are chosen, considering factors such as model complexity and computational efficiency. These features form a feature subset.
- An additional layer for *MI* feature selection is incorporated after the fusion of the two models. This layer selects suitable features from both models and combines them into a unified feature vector.
- The combined feature vector is integrated into the deep model, enhancing its feature representation for scene recognition tasks.

By following this *MI*-based feature selection algorithm, our model systematically identifies and integrates the most informative features from the first fully connected layer of the two models, effectively enhancing the discriminative power and accuracy of the feature set for precise and efficient scene recognition.

3.3. Dataset Used

In the pursuit of advancing scene recognition capabilities within the scope of our research, we employed two significant datasets to facilitate comprehensive model training and evaluation. We discussed the two datasets as follows.

3.3.1. Scene Classification Dataset

In our quest to advance scene recognition capabilities within our research, we leveraged the 'Scene Classification' dataset [28], an invaluable resource that significantly contributed to the diverse and comprehensive training and evaluation of our scene recognition model. This dataset is notable for its vast collection of approximately 25,000 images, which span a broad spectrum of natural scenes from various locations around the world. The primary objective of this dataset is to address the scene classification task, which involves the categorization of images into specific scene types based on their visual characteristics. In our specific study, the task involved a six-class classification problem, where each image was expected to be categorized into one of the following scene classes: Buildings, Forests, Mountains, Glacier, Street, or Sea.

To establish a consistent and standardized approach to dataset preparation, we meticulously selected and worked with 500 images from each of the six scene classes, resulting in a total of 3000 images for our study. These images were strategically partitioned into two essential subsets: training and testing. Approximately 80% of the images were dedicated to the training subset, allowing our model to learn and extract vital scene features. The remaining 20% of the images constituted the testing subset, serving as a critical benchmark to rigorously evaluate the model's scene recognition proficiency and generalization capabilities. In alignment with best practices for image dataset management, all images within the 'Scene Classification' dataset were uniformly resized to a resolution of 200×200 pixels. This resizing procedure not only facilitated computational efficiency but also ensured consistent image dimensions across the dataset, promoting a fair and equitable evaluation of the model's performance. The ground truth labeling of images in this dataset, which entails accurately associating each image with its respective scene class, was undertaken with utmost precision. This process ensured the reliability and trustworthiness of the scene labels, a critical factor underpinning the success of our scene recognition model. Figure 3 shows examples from this dataset.

Figure 3. Samples of images from scene classification dataset [28].

3.3.2. The AID Dataset

The second dataset employed in our study is the AID dataset [29], a valuable resource that significantly contributes to the diversification of scene recognition tasks. The AID dataset is characterized by its expansive collection, encompassing 30 distinct scene classes, each encapsulating the unique characteristics of various aerial landscapes. Within each of these scene classes, the dataset includes a substantial number of samples, ranging from 200 to 400 images, all uniformly sized at 600 × 600 pixels. To ensure methodical experimentation and robust model development, we strategically worked with a subset of 200 images from each of the 30 scene classes within the AID dataset. This subset facilitated both model training and evaluation processes, with an 80% allocation for training and a 20% allocation for testing, adhering to best practices in dataset partitioning.

What sets the AID dataset apart is its origin—comprising sample images meticulously collected from Google Earth imagery. While these images undergo post-processing through RGB renderings derived from the original optical aerial images, it has been extensively validated that there is no significant disparity between Google Earth images and genuine optical aerial images. This validation extends even to pixel-level land use and cover mapping, emphasizing the suitability of Google Earth imagery for evaluating scene classification algorithms. The AID dataset encompasses an extensive array of aerial scene types, including but not limited to airports, bridges, forests, meadows, schools, stadiums, and more. Notably, each image within the dataset has been expertly annotated by specialists in the field of remote sensing image interpretation, ensuring the precision and reliability of scene labels. These labels play a crucial role in training and evaluating scene recognition models, contributing to the dataset's overall utility.

One of the dataset's distinguishing features is its multisource nature, as it amalgamates images from various remote imaging sensors present in Google Earth. This characteristic presents additional challenges for scene classification compared to datasets containing single-source images. Moreover, the AID dataset's inclusivity extends across different countries and regions globally, with a diverse collection of images extracted under varying timeframes, seasons, and imaging conditions. This diversity enhances the dataset's intrinsic intra-class variations, further enriching the data's representativeness and applicability in scene recognition research. Figure 4 shows samples of images from this dataset.

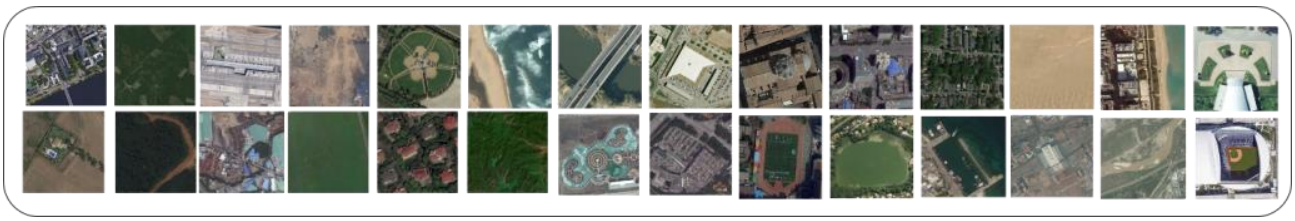

Figure 4. Samples of images from AID dataset [29].

3.4. Scene Recognition

The recognition stage in scene recognition is the culmination of the entire process, where the model employs the extracted and refined features to make informed decisions about the category or label of a given scene or image. This stage plays a pivotal role in the overall success of a scene recognition system, and its effectiveness directly hinges on the quality of the features obtained through prior processing, including feature selection and fusion. After feature selection, the recognition stage begins by taking the optimized feature representation as the input. These features are intended to capture the most discriminative and relevant aspects of the scene, having undergone a rigorous selection process to eliminate noise and redundancy. The choice of features is critical, as it greatly influences the model's ability to discern subtle differences between scenes.

Typically, the recognition stage is characterized by the presence of a SoftMax layer, which is the final layer of our deep model. The SoftMax layer transforms the output of the previous layers into probability distributions over the possible scene categories. Each output neuron corresponds to a specific scene class, and the SoftMax function ensures that the class probabilities sum up to one. Consequently, the class with the highest probability is assigned as the predicted label for the input scene.

To train the recognition model, a suitable loss function is employed, such as categorical cross-entropy, which quantifies the dissimilarity between predicted and ground truth labels. During training, the model learns to adjust its parameters (weights and biases) to minimize this loss, improving its ability to make accurate predictions. The model's performance is rigorously evaluated during the recognition stage using a separate validation or test dataset. Metrics such as accuracy, precision, recall, and *F1* score are commonly used to assess the model's ability to correctly classify scenes into their respective categories.

4. Results and Analysis

In this section, we present the results and in-depth analysis of our proposed method for scene recognition. The experiments were conducted using MATLAB R2019b with the deep learning toolbox on a machine equipped with an Intel Core i5 CPU operating at 2.4 GHz, 24 MB RAM, and an NVIDIA GeForce GTX 1650 GPU with 16 MB memory. Our method involved novel approaches such as multimodal deep learning, feature selection, and fusion. We employed two distinct datasets: the scene classification dataset [28] and the AID dataset [29]. Each dataset's image dimensions are matched to its characteristics and needs. To balance accuracy and efficiency, a lower resolution is suitable for datasets with fewer classes and complexity. A greater resolution helps the model catch scene features in datasets with more different classes.

4.1. Experimental Setup

For each dataset, we divided the images into training and testing sets, adhering to the 80–20 partitioning scheme. This division enabled us to train our model on a substantial portion of the data while reserving a separate set for rigorous evaluation. Our deep learning model leveraged a carefully crafted architecture, as discussed earlier, involving convolutional layers, feature selection, and fusion mechanisms. The model was trained using the Adam optimizer with specific hyperparameters. We employed standard evaluation metrics, including accuracy, precision, recall, and *F1* score, to quantitatively assess the performance of our method. These metrics offer valuable insights into the model's capacity to accurately categorize situations and its overall efficacy. The metrics can be calculated as follows [30]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where *TP* (true positives) is the number of samples correctly classified as positive (belonging to the target class). *TN* (true negatives): The number of samples correctly classified as negative (not belonging to the target class). *FP* (false positives): The number of samples

incorrectly classified as positive. *FN* (false negatives): The number of samples incorrectly classified as negative.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

4.2. Results on Scene Classification Dataset

In our experimentation, we employed a dataset comprising a total of 3000 images, with 2400 images designated for training purposes and the remaining 600 images reserved for testing. To comprehensively assess the performance of our scene recognition model, we conducted several analyses and evaluations. Figure 5 shows the accuracy and loss curves, which provide critical insights into the training and validation performance of our scene recognition model. These curves, which illustrate the model's progress over epochs, serve as valuable tools for assessing its convergence, stability, and generalization capabilities.

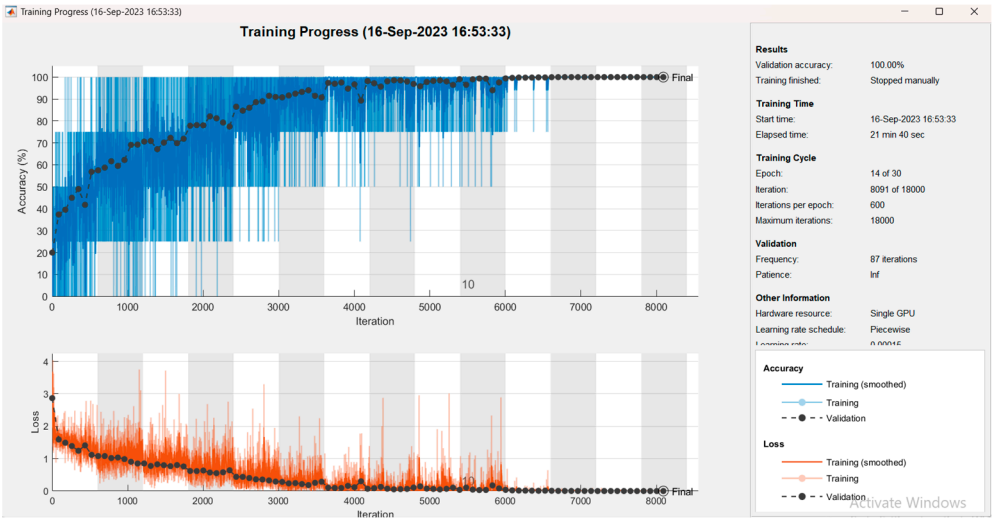

Figure 5. Training and validation progress of our method on the scene classification dataset.

In the figure, the accuracy curves for both training and validation showcase the model's ability to correctly classify scenes across various categories. From the outset, at iteration 0, both training and validation accuracies were observed at approximately 20%. As training progressed, these accuracies demonstrated steady and consistent improvement. By epoch 10, the accuracy curves had reached remarkable stability, with the validation accuracy peaking at 100%. This high level of validation accuracy signifies the model's robustness and proficiency in recognizing scenes. Notably, the curves remained stable through epoch 14, with no substantial changes observed. This stability prompted manual termination of the curve analysis at epoch 14, as no further significant improvements were anticipated.

The loss curves, indicative of the model's optimization progress, revealed a similar trend. At the inception of training, both training and validation losses began at approximately 3%. Subsequently, these losses gradually decreased, indicative of the model's capacity to minimize errors and enhance its scene recognition capabilities. By epoch 8, the loss curves had reached a state of stability, with both training and validation losses converging to 0%. This remarkable achievement underscores the model's effective learning and convergence capabilities, with minimal discrepancies between training and validation

losses. It is worth noting that this remarkable training and validation performance was achieved efficiently. The training process, from start to finish, required a total elapsed time of 21 min and 40 s. This efficiency in achieving high accuracy with minimal loss underscores the effectiveness and computational efficiency of our scene recognition model.

The confusion matrix presented in Figure 6 provides a detailed breakdown of the classification performance of our scene recognition model on the test sets across six distinct scene categories: Building, Forests, Glacier, Mountains, Sea, and Street. From the figure, the model achieved perfect classification, with all 101 samples correctly identified as such for the Building class. Similarly, the Forests class also yielded perfect classification, with all 103 samples correctly identified. The Glacier class exhibited impeccable performance, with all 124 samples correctly classified. For the Mountains class, the model achieved a high level of accuracy, with 83 out of 83 samples correctly classified. The Sea class was also accurately recognized, with all 102 samples correctly classified. In the Street class, the model displayed strong performance, with 87 out of 87 samples correctly classified.

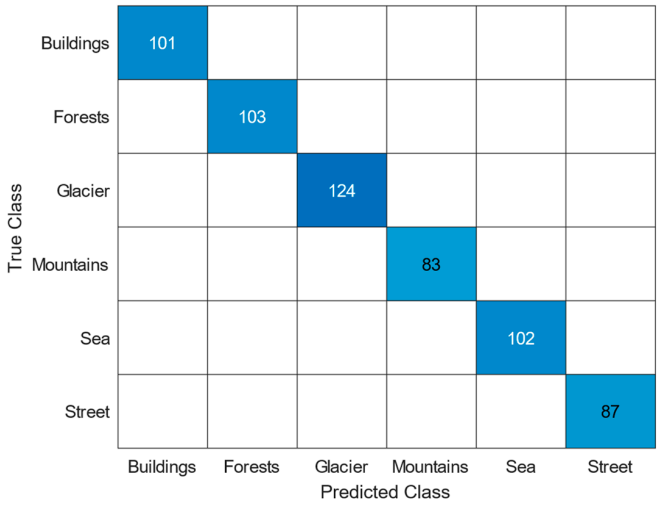

Figure 6. Confusion matrix for our model on the scene classification dataset.

4.3. Results on AID Dataset

In our experimentation, we employed a dataset comprising a total of 6000 images, with 4800 images designated for training purposes and the remaining 1200 images reserved for testing. Figure 7 shows the accuracy and loss curves for training and validation data from the AID dataset. The accuracy curves for both training and validation commenced at a modest 10% accuracy at the start of training, indicating a baseline level of recognition. As the training process unfolded, both curves exhibited a gradual yet consistent increase in accuracy. By the time epoch 11 was reached, the accuracy curves stabilized, with the validation accuracy reaching an impressive 98.02%. This high validation accuracy signifies the model's strong ability to accurately classify scenes across diverse categories. The stability observed in the curves persisted through epoch 13, with no significant fluctuations or improvements noted. Consequently, the decision was made to manually terminate the curve analysis at epoch 13 due to the sustained stability.

Parallel to the accuracy curves, the loss curves, which illustrate the model's optimization progress, displayed a similar pattern. At the outset, both training and validation losses began at approximately 3%, reflecting a starting point for error. Subsequently, these losses gradually decreased as the model refined its recognition capabilities. By the time epoch 7 was reached, both training and validation losses had converted to a minimal value of 0%. This convergence underscores the model's effective learning process and its ability to minimize errors effectively. It is noteworthy that the remarkable training and validation performance was achieved efficiently, with the entire training process taking 149 min and 48 s. This efficiency further underscores the practical feasibility of our scene recognition

model for real-world applications, as it combines both high accuracy and minimal loss in a relatively short time frame.

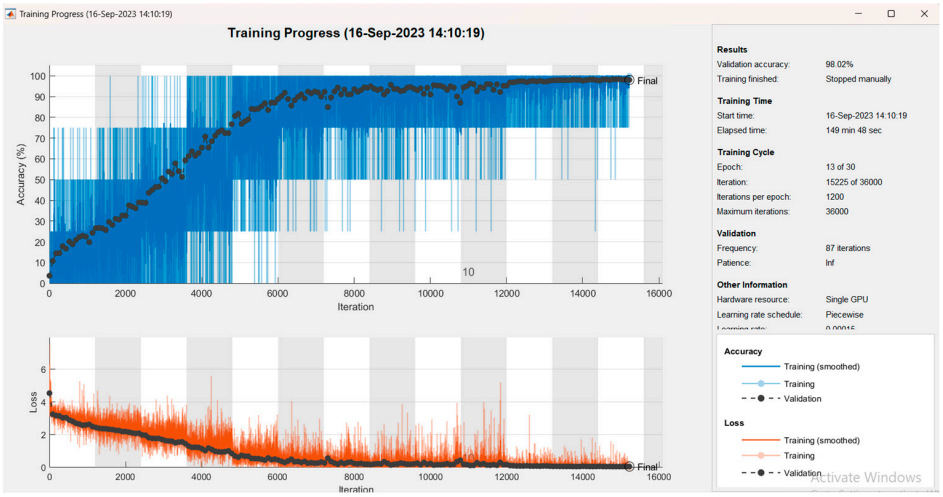

Figure 7. Training and validation progress of our method on the AID dataset.

Figure 8 shows the confusion matrix of the proposed method on the test sets of the second database. From the figure, we can observe that the Airport and Beach classes displayed a relatively high level of misclassification, with 27 out of 28 images for the Airport class and 13 out of 14 images correctly identified from the Beach class. However, one sample was incorrectly classified as Beach for the Airport class and was incorrectly classified as BaseballField for the Beach class. Further investigation may be required to address this misclassification. The Center class exhibited accurate classification for most samples, with 26 out of 28 samples correctly identified. Nevertheless, one sample was misclassified as BareLand and one sample was incorrectly classified as Beach. In the Church class, the model demonstrated strong performance, with 52 out of 54 samples correctly classified. Two samples, however, were misclassified as Desert and FarmLand. The Commercial class achieved excellent results, with 62 out of 64 samples correctly identified. Two samples were incorrectly classified as Meadow. Also, the Forest class achieved a good result with only two samples being incorrectly identified as Mountain. In the Parking and River classes, most samples, 61 out of 62 for Parking and 19 out of 20 for River, were correctly classified. One sample, however, was misclassified as Storage Tanks in both classes. The remaining classes were all classified correctly without any misclassifications.

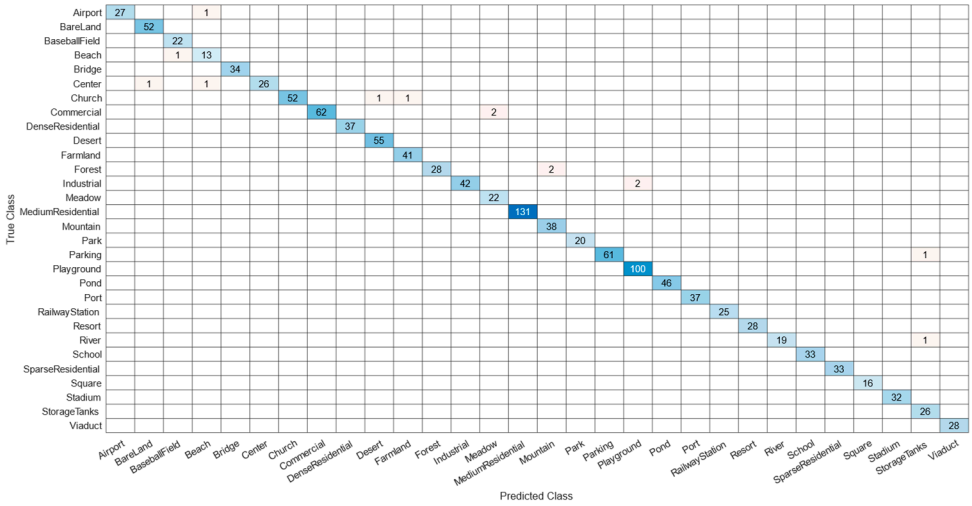

Figure 8. Confusion matrix for our model on the AID dataset.

5. Discussion

The results of our scene recognition model demonstrate its remarkable performance in accurately classifying scenes across multiple diverse categories. The core of our methodology revolves around the innovative concept of integrating multiple modalities into a single input for scene recognition. This approach significantly deviates from traditional models that predominantly rely on processing data from a single modality. To further optimize the feature representations resulting from the multimodal fusion, we introduced the technique of *MI*-based feature selection. This novel approach improves the quality, discriminative capability, and resilience of the features, thus providing significant advantages for scene identification tasks. This innovation is a key contribution to our model, as it enhances the effectiveness of feature representations. In the first experiment conducted on the Scene Classification dataset, which consists of six scene classes (Buildings, Forests, Glacier, Mountains, Sea, and Street), our model achieved outstanding results. In the second experiment conducted on the AID dataset, which consists of 30 different scene classes, our model also achieved good results compared to other methods using the same dataset. Table 3 shows a summary of the total performance of our method on the two datasets. These results illustrate the model's capacity to adapt and effectively handle a diverse array of scenario categories.

Table 3. Overall performance of our method on the two datasets.

Database	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Scene Classification [28]	100	100	100	100
AID [29]	98.83	98.83	98.83	98.83

From Figures 5 and 7, the accuracy and loss curves depict the successful training and validation of our scene recognition model. The stability achieved in both accuracy and loss, along with the rapid convergence to high accuracy and minimal loss values, highlights the effectiveness of our approach in accurately classifying scenes. Moreover, the efficient training process underscores the practical feasibility of our model for real-world applications. From Figures 6 and 8, the confusion matrices underscore the exceptional classification performance of our scene recognition model. The absence of misclassifications within the matrix highlights the model's precision and effectiveness in identifying diverse scenes, making it a robust and reliable tool for scene recognition tasks.

In comparison to previous studies in the field of scene recognition, our model stands out due to its multimodal approach and the innovative feature selection technique shown in Table 4. Traditional unimodal methods have been surpassed by our approach, which leverages the power of multimodal fusion directly on the same input image data. This advancement is crucial for applications such as autonomous vehicles, intelligent surveillance systems, and augmented reality, all of which rely on comprehensive scene understanding and interpretation.

From the table, we can observe that our approach demonstrates superior performance compared with other previous methods for scene recognition. Compared with the method presented by Zhao et al. [13], our accuracy is slightly lower at 98.83%, but it is still competitive, considering the exceptional performance of their approach. Moreover, our method may offer advantages in terms of complexity and resource requirements. Finally, our approach's ability to combine CNN with *MI* contributes to its superior performance, making it a promising solution for scene recognition tasks. We aim to highlight that the extraordinary performance of our model may be attributed to our creative approach, rigorous methodology, and the integration of feature selection based on *MI*. Our model has been meticulously crafted to possess both robustness and efficacy in the classification of situations over a wide range of categories.

Table 4. Comparison of our model with other previous models.

Ref./Author	Methods	Dataset	Performance
Hua et al. [10]	Multi-scene recognition network	AID	Precision = 64.03% Recall = 52.30% F1 Score = 52.39%
Petrovska et al. [11]	Pre-trained deep models + SVM	AID	Overall accuracy = 93.58%
Wang et al. [12]	CSDS model	AID	Overall accuracy = 94.29%
Zhao et al. [13]	Residual dense network	AID	Accuracy = 99%
Bazi et al. [14]	Vision Transformers	AID	Best Accuracy = 95.51%
Wang and Yu [15]	Deep learning	AID	Mean Accuracy = 93.70%
Wu et al. [16]	SDM + MAT	AID	Accuracy = 90.90%
Lima and Marfurt [17]	CNN	AID	Accuracy = 94.10%
Our	CNN + *MI*	AID	Accuracy = 98.83% Precision = 98.83% Recall = 98.83% F1 Score = 98.83%

While our model has demonstrated remarkable performance, it is essential to acknowledge its limitations. Future research could focus on further enhancing the model's ability to handle more extensive datasets and complex scenes. Additionally, efforts can be directed towards increasing the model's efficiency and reducing computational requirements, making it more accessible for real-time applications.

6. Conclusions

Our proposed scene recognition model integrates a CNN with *MI*-based feature selection, representing a significant advancement in scene recognition. By leveraging the power of multimodal deep learning and incorporating novel feature selection techniques, our method achieves impressive accuracies of 100% and 98.83% on the scene classification dataset and AID dataset, respectively. This surpasses the performance of several existing methods and demonstrates our model's robustness and precision in classifying scenes accurately. Moreover, our approach offers the advantage of reducing complexity and resource requirements compared to some high-performing alternatives. By combining the strengths of CNN- and *MI*-based feature selection, our method not only enhances scene recognition accuracy but also provides a foundation for further advancements in scene understanding and interpretation. This work contributes to the growing body of research aimed at enabling machines to comprehend and navigate complex visual environments, with potential applications in autonomous vehicles, surveillance systems, augmented reality, and more.

Author Contributions: Conceptualization, S.A.C. and W.A.; methodology, M.H. and A.A.A.E.-L.; software, M.H.; validation, M.H., S.A.C., and A.A.A.E.-L.; formal analysis, W.A. and A.A.A.E.-L.; investigation, M.H.; resources, S.A.C. and W.A.; data curation, M.H. and A.A.A.E.-L.; writing—original draft preparation, M.H., S.A.C., and A.A.A.E.-L.; writing—review and editing, M.H. and A.A.A.E.-L.; visualization, M.H. and W.A.; supervision, A.A.A.E.-L.; project administration, S.A.C. and W.A.; funding acquisition, S.A.C. and W.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education, in Saudi Arabia for funding this research work through the project number RI-44-0490.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available at https://www.kaggle.com/datasets/nitishabharathi/scene-classification (accessed on 6 October 2023) and https://www.kaggle.com/datasets/jiayuanchengala/aid-scene-classification-datasets (accessed on 6 October 2023).

Acknowledgments: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education, in Saudi Arabia for funding this research work through the project number RI-44-0490.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xie, L.; Lee, F.; Liu, L.; Kotani, K.; Chen, Q. Scene recognition: A comprehensive survey. *Pattern Recognit.* **2020**, *102*, 107205.
2. Moeslund, T.B.; Hilton, A.; Krüger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **2006**, *104*, 90–126.
3. Gupta, A.; Anpalagan, A.; Guan, L.; Khwaja, A.S. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array* **2021**, *10*, 100057.
4. Zhang, J.; Zhu, C.; Zheng, L.; Xu, K. ROSEFusion: Random optimization for online dense reconstruction under fast camera motion. *ACM Trans. Graph. TOG* **2021**, *40*, 1–17.
5. Saber, S.; Amin, K.; Pławiak, P.; Tadeusiewicz, R.; Hammad, M. Graph convolutional network with triplet attention learning for person re-identification. *Inf. Sci.* **2022**, *617*, 331–345.
6. Saber, S.; Meshoul, S.; Amin, K.; Pławiak, P.; Hammad, M. A Multi-Attention Approach for Person Re-Identification Using Deep Learning. *Sensors* **2023**, *23*, 3678.
7. Guan, T.; Wang, C.H. Registration based on scene recognition and natural features tracking techniques for wide-area augmented reality systems. *IEEE Trans. Multimed.* **2009**, *11*, 1393–1406.
8. Pawar, P.G.; Devendran, V. Scene understanding: A survey to see the world at a single glance. In Proceedings of the 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 28–29 September 2019; IEEE: New York, NY, USA; pp. 182–186.
9. Huang, N.; Liu, Y.; Zhang, Q.; Han, J. Joint cross-modal and unimodal features for RGB-D salient object detection. *IEEE Trans. Multimed.* **2020**, *23*, 2428–2441.
10. Hua, Y.; Mou, L.; Lin, J.; Heidler, K.; Zhu, X.X. Aerial scene understanding in the wild: Multi-scene recognition via prototype-based memory networks. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 89–102.
11. Petrovska, B.; Atanasova-Pacemska, T.; Corizzo, R.; Mignone, P.; Lameski, P.; Zdravevski, E. Aerial scene classification through fine-tuning with adaptive learning rates and label smoothing. *Appl. Sci.* **2020**, *10*, 5792.
12. Wang, X.; Yuan, L.; Xu, H.; Wen, X. CSDS: End-to-end aerial scenes classification with depthwise separable convolution and an attention mechanism. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10484–10499.
13. Zhao, X.; Zhang, J.; Tian, J.; Zhuo, L.; Zhang, J. Residual dense network based on channel-spatial attention for the scene classification of a high-resolution remote sensing image. *Remote Sens.* **2020**, *12*, 1887.
14. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision transformers for remote sensing image classification. *Remote Sens.* **2021**, *13*, 516.
15. Wang, H.; Yu, Y. Deep feature fusion for high-resolution aerial scene classification. *Neural Process. Lett.* **2020**, *51*, 853–865.
16. Wu, H.; Xu, C.; Liu, H. S-MAT: Semantic-Driven Masked Attention Transformer for Multi-Label Aerial Image Classification. *Sensors* **2022**, *22*, 5433.
17. Pires de Lima, R.; Marfurt, K. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sens.* **2019**, *12*, 86.
18. Sharma, T.; Debaque, B.; Duclos, N.; Chehri, A.; Kinder, B.; Fortier, P. Deep learning-based object detection and scene perception under bad weather conditions. *Electronics* **2022**, *11*, 563.
19. Wang, S.; Yao, S.; Niu, K.; Dong, C.; Qin, C.; Zhuang, H. Intelligent scene recognition based on deep learning. *IEEE Access* **2021**, *9*, 24984–24993.
20. Afif, M.; Ayachi, R.; Said, Y.; Atri, M. Deep learning based application for indoor scene recognition. *Neural Process. Lett.* **2020**, *51*, 2827–2837.
21. Dhanaraj, M.; Sharma, M.; Sarkar, T.; Karnam, S.; Chachlakis, D.; Ptucha, R.; Markopoulos, P.P.; Saber, E. Vehicle detection from multi-modal aerial imagery using YOLOv3 with mid-level fusion. In Proceedings of the Big Data II: Learning, Analytics, and Applications, Online, 15 May 2020; SPIE: New York, NY, USA; Volume 11395, pp. 22–32.
22. Shahzad, H.M.; Bhatti, S.M.; Jaffar, A.; Rashid, M.; Akram, S. Multi-Modal CNN Features Fusion for Emotion Recognition: A Modified Xception Model. *IEEE Access* **2023**, *11*, 94281–94289.
23. Xu, H.; Huang, C.; Huang, X.; Huang, M. Multi-modal multi-concept-based deep neural network for automatic image annotation. *Multimed. Tools Appl.* **2019**, *78*, 30651–30675.

24. Doquire, G.; Verleysen, M. Mutual information-based feature selection for multilabel classification. *Neurocomputing* **2013**, *122*, 148–155.
25. Hu, Q.; Zhang, L.; Zhang, D.; Pan, W.; An, S.; Pedrycz, W. Measuring relevance between discrete and continuous features based on neighborhood mutual information. *Expert Syst. Appl.* **2011**, *38*, 10737–10750.
26. Liu, X.; Wang, S.; Lu, S.; Yin, Z.; Li, X.; Yin, L.; Tian, J.; Zheng, W. Adapting Feature Selection Algorithms for the Classification of Chinese Texts. *Systems* **2023**, *11*, 483.
27. Lu, S.; Ding, Y.; Liu, M.; Yin, Z.; Yin, L.; Zheng, W. Multiscale feature extraction and fusion of image and text in VQA. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 54.
28. Nitisha. Scene Classification. 2018. Available online: https://www.kaggle.com/datasets/nitishabharathi/scene-classification (accessed on 17 September 2023).
29. JayChen. AID: A Scene Classification Dataset. 2022. Available online: https://www.kaggle.com/datasets/jiayuanchengala/aid-scene-classification-datasets (accessed on 17 September 2023).
30. Manning, C.D. *An Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2009.