

Competent scene classification using feature fusion of pre-trained convolutional neural networks

Thirumaladevi Satharajupalli¹, Kilari Veera Swamy², Maruvada Sailaja¹

¹ECE Department, Jawaharlal Nehru Technological University, Kakinada-533003, Andhra Pradesh, India

²ECE Department, Vasavi College of Engineering, Ibrahimbagh, Hyderabad-500 031, Telangana, India

Article Info

Article history:

Received Aug 29, 2022

Revised Dec 10, 2022

Accepted Feb 16, 2023

Keywords:

Feature extraction

Feature fusion

Pre-trained networks

Scene classification

Support vector machine

ABSTRACT

In view of the fact that the development of convolutional neural networks (CNN) and other deep learning techniques, scientists have become more interested in the scene categorization of remotely acquired images as well as other algorithms and datasets. The spatial geometric detail information may be lost as the convolution layer thickness increases, which would have a significant impact on the classification accuracy. Fusion-based techniques, which are regarded to be a viable way to express scene features, have recently attracted a lot of interest as a solution to this issue. Here, we suggested a convolutional feature fusion network that makes use of canonical correlation, which is the linear correlation between two feature maps. Then, to improve scene classification accuracy, the deep features extracted from various pre-trained convolutional neural networks are efficiently fused. We thoroughly evaluated three different fused CNN designs to achieve the best results. Finally, we used the support vector machine for categorization (SVM). In the analysis, two real-world datasets UC Merced and SIRI-WHU were employed, and the competitiveness of the investigated technique was evaluated. The improved categorization accuracy demonstrates that the fusion technique under consideration has produced affirmative results when compared to individual networks.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Thirumaladevi Satharajupalli

ECE Department, Jawaharlal Nehru Technological University

Kakinada-533003, Andhra Pradesh, India

Email: thirumaladevice@gmail.com

1. INTRODUCTION

Classifying remote sensing images into various classes depending on image content has gotten a lot of attention nowadays because of its wide range of applications. Classification of remote sensing image scenes is primarily a machine learning and computer vision problem. The investigation of scene classification entails convolutional neural networks (CNNs) were highly successful. However, to train their parameter sets, most of these models require vast amounts of labeled data and many iterations. Several CNN-based scene categorization approaches [1]-[3] have arisen as a result of varied tactics for utilizing CNNs. Deep learning models, such as AlexNet [4], and VGG-Net [5], have achieved considerable success in computer training data [6], ImageNet [7]. In 2012, Krizhevsky took first place in the ImageNet large scale visual recognition challenge. For a specific purpose, acquiring a large dataset can be expensive. When it comes to remote sensing image scene categorization, deep neural network model-based methods are becoming increasingly popular [8], [9]. Deep learning-based scene image categorization has a distinct advantage over typical machine learning methods in that can be extracted from more complicated and relevant feature structures [10], [11]. Deep layers can then be used to get crucial and discriminatory feature representation, while irrelevant versions are ignored.

There are three types of CNN-based techniques utilizing pre-trained utilizing CNNs to act as feature extractors and fine-tuning already trained models on target datasets and creating new CNN models for scene categorization from scratch [12], [13]. The way that CNN is affianced as a feature extractor is the simplest of the three approaches. CNNs were first introduced as feature extractors. In 2015, Penatti *et al.* [14] applied CNNs to image scene categorization using remote sensing and investigated the adaptive ability of off-the-shelf CNNs for remote sensing image categorization. CNNs outperform low-level descriptors in their studies, according to the researchers. Later, Cheng *et al.* [15] researched how to fully utilize pre-trained CNNs for scene categorization by treating them as feature extractors. Marmaris *et al.* [16] presented a two-stage CNN scene categorization framework. It utilized CNNs that had been pre-trained to extract a collection of interpretations from images. Covariance-based multilayer feature fusion is also proposed [17]. Classifiers were then given the extracted representations.

Various structures of CNNs have varying receptive fields and may capture different types of data from images. In this paper, AlexNet, VGG-19, and VGG-16 are employed as deep feature extractors independently, and to improve accuracy, features extracted by two of three are fused using canonical correlation analysis (CCA) and form three fused networks, which are assessed and compared. The datasets from UC Merced and SIRI-WHU, both of which are publicly available and created for research purposes, were used to assess the performance of the fused networks.

The remainder of this work is structured in a succeeding manner. Section 2 shows and explains the scene classification workflow established on the individual pre-trained CNN model including a flow chart of three fusion networks. In section 3, we will find typical datasets that exhibit experimental evaluation and analysis. This paper comes to an end with the section 4 conclusion.

2. WORKFLOW OF THE PROPOSED METHOD

Convolutional, pooling, and fully connected layer are the core components of the pre-trained CNN model, which only use pre-trained weights [18]. Figure 1 shows a distinctive remote sensing image scene categorization procedure using a CNN model that has been pre-trained. Samples of the input training images are loaded into the input layer, which pre-processes and modifies the image dimensions as per the network model's input layer. The finishing output is achieved with a fully connected layer after some a succession of pooling and convolutional computations which are then used as final features and applied to support vector machine (SVM) for classification.

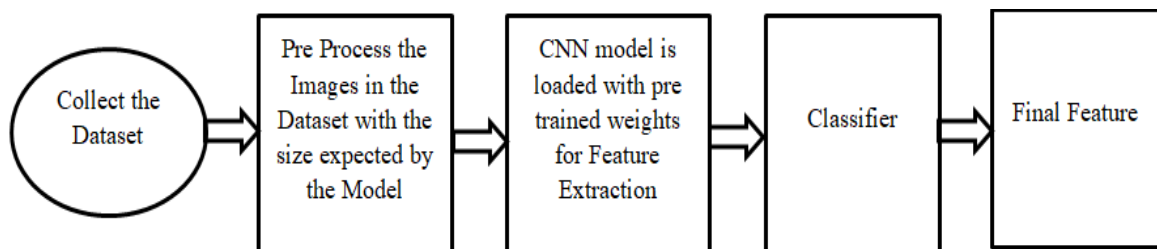


Figure 1. A feature extractor's flowchart that makes use of pre-trained CNN

For image scene classification, we employed pre-trained deep CNN models, where used AlexNet, VGG-19, and VGG-16 as feature extractors and selected helpful layers to acquire a good depiction of the image scene. We are the first to integrate several fully connected layers of the AlexNet, VGG-Net architecture, where each layer's output is expected to be a feature descriptor and is fused to produce a final feature illustration of the input image. Other individual feature representation methods perform less than fused deep feature learning. The visual scene is well described by fused features, which offer a lot of information.

In very high resolution (VHR) image scene categorization, the purpose of feature fusion is to combine two correlated scene features into a single feature vector with much more discriminant data than the input feature vectors. Here novelty of this paper is combining characteristics collected from dense layers of two different pre-trained networks to create various fusion networks. Finally, a multi-kernel learning approach was used to train the SVM classifier, as illustrated in Figure 2.

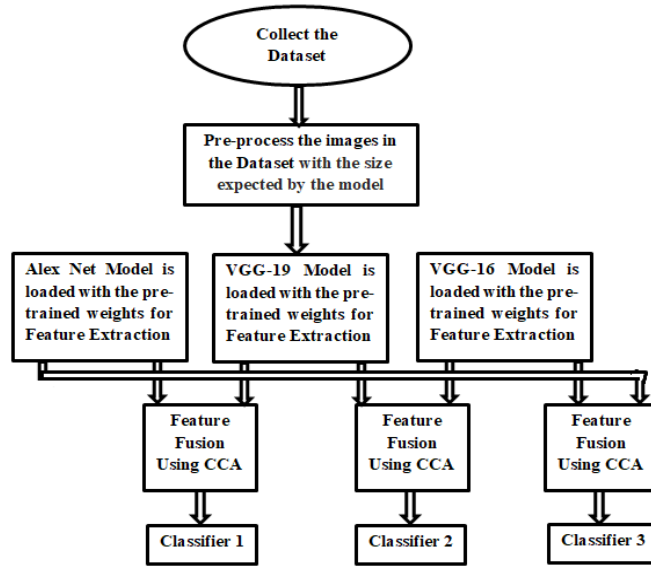


Figure 2. Proposed fused pre-trained CNNs flowchart for scene categorization

Combining two or more characteristics properly is becoming increasingly difficult. We are using the canonical correlation approach to combine two distinct feature vectors to create a new feature vector that is significantly more discriminative than the original two. This approach transforms a correlation analysis of two random vectors into a few uncorrelated pairs of variables. M and N are two zero-mean arbitrary vector matrices. Determine a couple of orientations that will improve the projection's correlation. Choose a collection of canonical variables. Two features matrices M and N , in addition, $M \in R^{axn}$ and $N \in R^{b \times n}$ where n is the training feature vectors containing the matrices. M 's covariance matrix is epitomized as $C_{pp} \in R^{axa}$ as well as $C_{qq} \in R^{b \times b}$ is the covariance matrix of N between-sets the covariance matrix is $C_{pq} \in R^{axb}$ and $C_{qp} = C_{pq}^T$. Overall covariance matrix $C \in R^{(a+b) \times (a+b)}$ is deliberate as:

$$C = \begin{pmatrix} var(p) & cov(p, q) \\ cov(q, p) & var(q) \end{pmatrix} = \begin{pmatrix} C_{pp} & C_{pq} \\ C_{qp} & C_{qq} \end{pmatrix} \quad (1)$$

It's difficult to deduce the correlations between all these matrices two pairs of feature vectors since the correlations between two groups of feature vectors might not conform to a predictable pattern. CCA's goal is to describe a linear combination [19] that maximizes correlation.

$$M^* = W_p^T M \text{ and } N^* = W_q^T N \quad (2)$$

$$Corr(M^*, N^*) = \frac{cov(M^*, N^*)}{var(M^*), var(N^*)} \quad (3)$$

Wherever:

- $(M^*) = W_p^T C_{pp} W_p$
- $var(N^*) = W_q^T C_{qq} W_q$
- $cov(M^*, N^*) = p^T C_{pq} W_q$

Exploiting the covariance between M^* and N^* considering the constraints $var(M^*) = var(N^*) = 1$. The transformation matrices W_p, W_q is initiated by especially the eigenvalue concerns.

$$\begin{aligned} C_{pp} - 1C_{pq} C_{qq} - 1C_{qp} W_p^{\wedge} &= \wedge^2 W_p^{\wedge} \\ C_{qq} - 1C_{qp} C_{pp} - 1C_{pq} W_q^{\wedge} &= \wedge^2 W_q^{\wedge} \end{aligned} \quad (4)$$

Where W_p^{\wedge} and W_q^{\wedge} are eigenvectors. \wedge^2 is the diagonal matrix with the highest eigenvalues or correlation squares. Each equation contains the following number of non-zero eigenvalues:

- $d = rank(C_{pq}) \leq \min(n, a, b)$ which is to be listed in descending manner $\alpha_1 \geq \alpha_2 \geq \alpha_3 \dots \geq \alpha_d$.
- $\alpha_1^T M$ and $\beta_1^T N$ (the first pair)
- $\alpha_2^T M$ and $\beta_2^T N$ (the second pair)
- $\alpha_d^T M$ and $\beta_d^T N$ (the d th pair)

$$\begin{aligned} M^* &= (\alpha_1^T M, \alpha_2^T M, \dots, \alpha_d^T M) = (\alpha_1, \alpha_2, \dots, \alpha_d)^T M = W_p^T \\ N^* &= (\beta_1^T N, \beta_2^T N, \dots, \beta_d^T N) = (\beta_1, \beta_2, \dots, \beta_d)^T N = W_q^T \end{aligned} \quad (5)$$

The ordered eigenvectors analogous to non-zero eigenvalues make up the transformation matrices W_p and W_q . A summing of the changed feature vectors is used for feature-level fusion. Canonical discriminant correlation The following features are included:

$$Z = M^* + N^* = (W_p^T M + W_q^T N) = \begin{pmatrix} W_p \\ W_q \end{pmatrix}^T \begin{pmatrix} M \\ N \end{pmatrix} \quad (6)$$

SVM: this categorization is based on the data of choice, which manages space with high-dimensional features with choice constraints. Using a set of labeled training datasets, SVM can produce linear capacity in either input space or maximum space [20]. This allows us to distinguish between positive and negative samples. A data matrix that has been labeled to either a positive or negative class is utilized as input data used in the SVM's training phase. One can utilize trained SVM to forecast what the class has predicted in test samples.

The benefits of CCA are straightforward to apply to two variables. The intermodality relationship is thought to be linear in CCA, and both modalities are interchangeable and given the same considerations. Linear feature transforms do not affect canonical correlations.

3. RESULTS AND DISCUSSION

3.1. UC Merced land-use dataset

UC Merced land-use dataset [21]: a lot of work has gone into making datasets available to the general public, including the first publicly accessible high-resolution remote sensing imagery collected for scene classification. This University of California Merced (UCM) land-use dataset [22] comprises 2,100 aerial shots of scenes distributed into 21 land-use scenario groups. Every single class contains 100 images that are 256×256 pixels in dimensions and have a resolution of every pixel, 0.3 meters. This dataset was created exhausting the United States Geological Survey's National Map, which was retrieved using aerial ortho imagery (USGS). This dataset features overlying land-use classifications, such as Figure 3 exhibits representations of sample images from every class Figure 3(a) agricultural, Figure 3(b) airplane, Figure 3(c) baseball diamond, Figure 3(d) beach, Figure 3(e) buildings, Figure 3(f) chaparral, Figure 3(g) denseresidential, Figure 3(h) forest, Figure 3(i) freeway, Figure 3(j) golfcourse, Figure 3(k) harbour, Figure 3(l) intersection, Figure 3(m) mediumresidential, Figure 3(n) mobilehomepark, Figure 3(o) overpass, Figure 3(p) parkinglot, Figure 3(q) river, Figure 3(r) runway, Figure 3(s) sparse residential, Figure 3(t) storagetanks and Figure 3(u) tennis court and it's been widely utilized for visual scene categorization and retrieval using remote sensing data.

3.2. SIRI-WHU

SIRI-WHU [23]: it's a collection of 2,400 remote sensing images that have been categorized into 12 scene types. Every single class comprises 200 images that are 200×200 pixels in size and have a resolution of 2 meters. It was sent to Wuhan University's intelligent data extraction and remote sensing (RS IDEA) group via Google Earth (Google Inc.). In Figure 4, sample images from the SIRI-WHU dataset are displayed Figure 4(a) agriculture, Figure 4(b) commercial, Figure 4(c) harbor, Figure 4(d) idle land, Figure 4(e) industrial, Figure 4(f) meadow, Figure 4(g) overpass, Figure 4(h) park, Figure 4(i) pond, Figure 4(j) residential, Figure 4(k) river, and Figure 4(l) water are among the 12 land-use categories. Even though this dataset has been studied using a variety of approaches [24], the number of scene classes is very small. It also concentrates mostly on China's urban areas.

3.3. AlexNet

AlexNet: the ImageNet large-scale image recognition competition (ILSVRC) was won by AlexNet [4] in 2012. The first to notice it was Alex Krizhevsky and his coworkers. This model has three fully connected layers in addition to a pooling layer and five convolutional layers. The initial and second convolutional layers are constrained by two normalization layers. This model accepts images with a 227×227 input size. The second fully-connected layer is where the 4096-pixel-long output feature vector is generated. SVM was used in our study to categorize the AlexNet CNN feature.



Figure 3. Example images of 21 classes' depiction of UC Merced dataset: (a) agricultural, (b) airplane, (c) baseball diamond, (d) beach, (e) buildings, (f) chaparral, (g) denseresidential, (h) forest, (i) freeway, (j) golfcourse, (k) harbou, (l) intersection, (m) mediumresidential, (n) mobilehomepark, (o) overpass, (p) parkinglot, (q) river, (r) runway, (s) sparse residential, (t) storagetanks, and (u) tenniscourt

3.4. VGG-Net

VGG-Net: a proposal was made in [5] and emerged victorious in the ILSVRC-2014 competition's localization and classification tasks. Two well-known architectures are VGG-19 and VGG-16. In this evaluation, the designs and overall performance were improved to a somewhat greater extent. This model contains 3 fully-connected layers, 5 pooling layers, and 13 convolutional layers. Images with a size of 224×224 are used as input to this model. The output feature vector is 4096 pixels long and derives from the second fully connected layer. In our evaluation, we employed SVM to classify the VGG-Net CNN feature.

3.5. Fused network

Fused network: two pre-trained networks are used to create this network. The input layers of the pre-trained networks continue to stay the same for AlexNet 227×227 , and for VGG-19, VGG-16 224×224 , up to feature extraction, the procedure is the same as for individual pre-trained networks, and the output feature is obtained from the second dense layer with dimension 4096, the fusion of these features can be done using the conical correlation concept already explained in section 2. It is possible to combine these two feature vectors from two distinct networks to obtain a final feature vector with a dimension of 4096. In our evaluation, we employed SVM to classify the fused net feature.

We employ the UC Merced dataset and the SIRI-WHU dataset to examine the scene categorization performance of individual pre-trained, fused networks. We use the experiment setup from [25] for evaluation, which chooses a total of 80 images from each category that will be used for training and the rest for testing. Regarding classification accuracy, the technique when fusion is absent is compared to several fusion methods, with fusion networks showing a significant improvement.



Figure 4. Example image depiction of the SIRI-WHU dataset: (a) agriculture, (b) commercial, (c) harbor, (d) idle land, (e) industrial, (f) meadow, (g) overpass, (h) park, (i) pond, (j) residential, (k) river, and (l) water

3.6. Metrics used for evaluation

Metrics used for evaluation: when it comes to image classification, the average accuracy, overall accuracy, and confusion matrix are the three metrics that are most frequently employed for evaluation purposes. The number of samples that have been correctly recognized is calculated by dividing the total number of samples by the category to which they are assigned. This division is performed irrespective of the quality of the overall assessment. The average accuracy is determined by adding up the classification accuracy of each class, regardless of the number of samples that are included in each category. The overall accuracy value and the average accuracy value are identical this is because the count for each class in the dataset is the same. Because of this, we only used overall accuracy and confusion matrix criteria to judge the performance of the various classification algorithms used in this work. The confusion matrix and overall accuracy measurements were also looked into to ensure consistency. A total of five iterations of the study confirmed that an 80 to 20 split between training and testing yielded the best results. The formula for calculating the accuracy, which is expressed as a percentage of correct occurrences, is:

$$Accuracy = \frac{TP}{TP+FN+TN+FP} \quad (7)$$

The proportion of truly positive occurrences that are expected to occur in all positive situations is known as precision. The formula is as:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

The recall calculation equation determines the predicted percent of true positive samples.

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

An exhaustive mathematical calculation, the F1-score considers both accuracy and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (10)$$

Figure 5 illustrates together the confusion matrix in addition to the accuracy. In the first row, networks Figure 5(a) AlexNet, Figure 5(b) VGG-19, and Figure 5(c) VGG-16 are the proposed individual pre-trained CNN-based network classification is shown, while for scene categorization, the final feature extractor is a single fully-connected layer FC7. In the second row, a fusion-based proposed network confusion matrix with accuracy is shown Figure 5(d) AlexNet–VGG-19, Figure 5(e) AlexNet–VGG-16, and Figure 5(f) VGG-19–VGG-16. Pre-trained AlexNets had an overall accuracy (OA) of 79.76 percent, VGG-19 had an OA of 81.19 percent, and VGG-16 had an OA of 83.81 percent, according to experiments on the University of California Merced dataset, but fusion-based pre-trained networks had an OA of 89.28 percent, 90.23 percent, and 91.42 percent, respectively. The proposed approach provides optimal classification efficiency for the majority of classes in the case of UCM with an average gain of 8% in classification accuracy.

In the testing of the SIRI-WHU dataset, the confusion matrix is shown in Figure 6. To classify scenes, the fully connected layer FC7 is employed as the final feature extractor, as illustrated in the first row, Figure 6(a) AlexNet, Figure 6(b) VGG-19, Figure 6(c) VGG-16 and potential fusion-based classification models are displayed in the second row as Figure 6(d) AlexNet-VGG-19, Figure 6(e) AlexNet–VGG-16, and Figure 6(f) VGG-19–VGG-16. Whereas the accuracy of AlexNet’s preprocessed single layer is 86.52 percent, VGG-19 is 87.60 percent, and VGG-16 is 88.04 percent, the proposed technique enhances accuracy by 90.62 percent, 91.87 percent, and 92.91 percent, respectively.

Table 1 illustrates the corresponding performance assessment for the two datasets. The proposed scenario indicates an increase in OA. When multiple individual pre-trained and fusion-based learning networks are utilized. As demonstrated in Figure 7 correspondingly Figure 7(a) UCM dataset and the Figure 7(b) SIRI-WHU dataset the proposed technique achieves optimal Improves categorization performance in the majority of classes and overall accuracy by 4%.

Table 1. Evaluation results comparison for the individual proposed fused network accuracies for the two datasets UCM, SIRI-WHU

Method	Network used	UCM dataset with 80% training (overall accuracy %)	SIRI-WHU dataset with 80% training (overall accuracy %)
Pre-trained individual network FC7 as a feature extractor	AlexNet	79.76	86.45
	VGG-VD19	81.19	87.7
	VGG-VD16	83.81	88.12
Proposed fusion network using CCA	AlexNet – VGG-19	89.28	90.62
	AlexNet – VGG-16	90.23	91.87
	VGG-19 – VGG-16	91.66	92.91

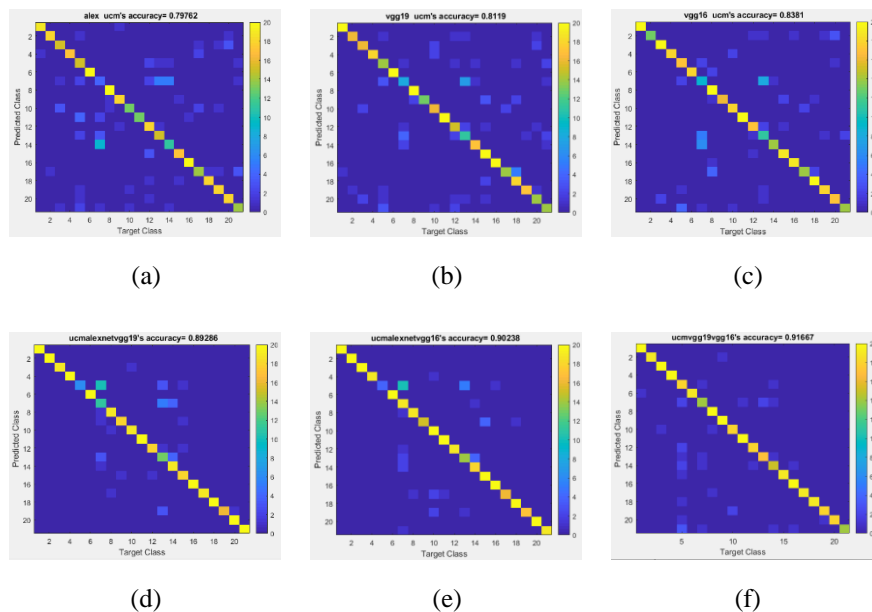


Figure 5. First row consistent to single-layered confusion matrix using three pre-trained networks: (a) AlexNet, (b) VGG-19, (c) VGG-16 second row resultant to proposed fusion, (d) AlexNet-VGG-19, (e) AlexNet-VGG16, and (f) VGG-19–VGG-16 of UC Merced dataset

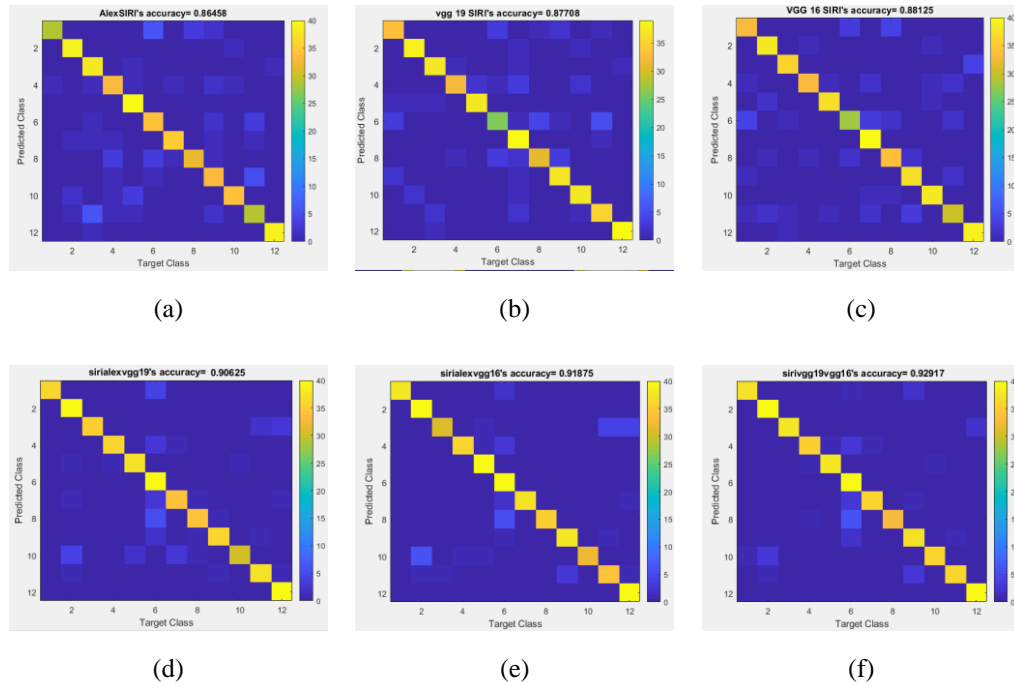


Figure 6. First row consistent to single-layered confusion matrix using three pre-trained networks: (a) AlexNet, (b) VGG-19, (c) VGG-16 second row resultant to proposed fusion, (d) AlexNet–VGG-19, (e) AlexNet–VGG-16, and (f) VGG-19–VGG-16 of SIRI-WHU dataset

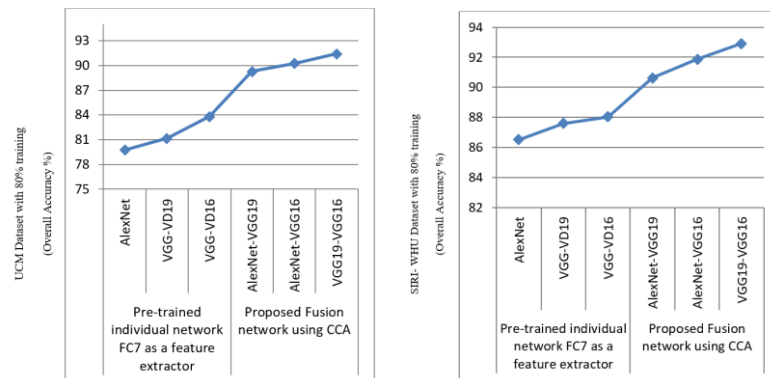


Figure 7. Proposed fusion-based networks versus pre-trained networks comparison of using (a) UC Merced and (b) SIRI-WHU datasets

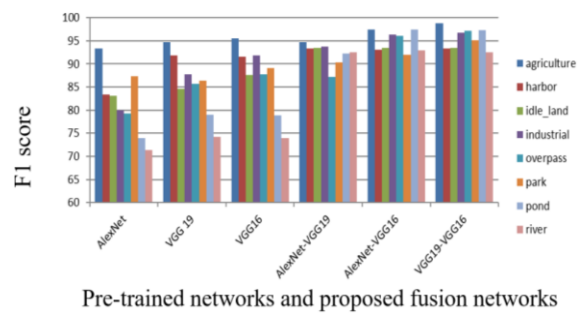
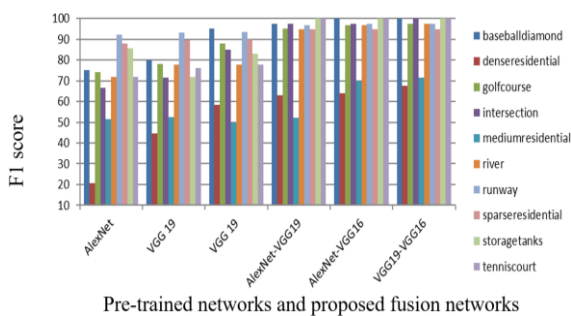


Figure 8. Comparison of F1 results when compared to pre-trained networks the UC Merced dataset improved classes using the suggested fusion networks

Figure 9. Comparison of F1 scores SIRI-WHU dataset improved classes using suggested fusion networks compared to ones that have already been pre-trained

From the UCM dataset, Figure 8 depicts the F1 scores. Individual pre-trained networks and proposed fusion-based approaches are presented with F1 scores of improved classes. The proposed fusion approach benefits a wide range of classes, including denseresidential, which improves accuracy from 20% to 68%, golfcourse, river, runway, sparseresidential, which improves accuracy from 90% to 100%, and baseballdiamond, intersection, storagetanks, and tennis court.

Figure 9 displays the SIRI-WHU dataset's recommended fusion networks and single-layered pre-trained networks F1 scores. As can be observed, the recommended strategy improves scores in the majority of classes. In the SIRI-WHU dataset, the river and pond improves from 70% to 95%. The classes industrial, overpass are achieved 80 to 90 % and majority of categories scored above 95%.

4. CONCLUSION

Significant advances in remote sensing technology have presented us with a torrent of remote sensing data for scene categorization using images from remote sensing throughout the previous decade. Because there is a scarcity of freely available remote-sensing image data, especially for deep learning-based expertise, which severely restricts the development of new approaches. The purpose of this research was to investigate in what manner machine learning also fusion network designs performed while categorizing data. Using the UCM and SIRI-WHU datasets, classification was done on three feasible architectures: AlexNet, VGG-19, and VGG-16. With accuracy from a pre-trained network of 89 percent, 90 percent, and 91 percent for the UCM dataset and 91 percent, 92 percent, and 93 percent for the SIRI-WHU dataset, the suggested methodology improved the state-of-the-art and created a standard. The classification results of the machine learning concept were compared using SVM. The use of fusion networks has been proven to be an effective strategy for optimal outcomes. This methodology can only determine the linear correlation between two features. Future extensions could include handling non-linear feature spaces with more than two dimensions.




REFERENCES

- [1] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, 2018, doi: 10.1109/TGRS.2017.2783902.
- [2] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for geospatial land classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6530–6541, 2019, doi: 10.1109/TGRS.2019.2906883.
- [3] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, 2019, doi: 10.1016/j.isprsjprs.2019.04.015.
- [4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.
- [5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR 2015*, 2015, pp. 1–14, doi: 10.48550/arXiv.1409.1556.
- [6] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [7] T. -Y. Lin *et al.*, "Microsoft COCO: common objects in context," in *European Conference on Computer Vision – ECCV 2014*, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.
- [8] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. F. -Fei, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
- [9] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1461–1474, 2020, doi: 10.1109/TNNLS.2019.2920374.
- [10] Sutikno, H. A. Wibawa, and P. S. Sasongko, "Detection of Ship using Image Processing and Neural Network," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 16, no. 1, pp. 259–264, 2018, doi: 10.12928/telkomnika.v16i1.7357.
- [11] Y. Yu, Z. Gong, C. Wang, and P. Zhong, "An unsupervised convolutional feature fusion network for deep representation of remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 1, pp. 23–27, 2018, doi: 10.1109/LGRS.2017.2767626.
- [12] B. K. O. C. Alwawi and A. F. Y. Althabhaee, "Towards more accurate and efficient human iris recognition model using deep learning technology," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 20, no. 4, pp. 817–824, 2022, doi: 10.12928/telkomnika.v20i4.23759.
- [13] Y. Liu, Y. Liu, and L. Ding, "Scene classification based on two-stage deep feature fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 183–186, 2018, doi: 10.1109/LGRS.2017.2779469.
- [14] O. A. B. Penatti, K. Nogueira, and J. A. D. Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 44–51, doi: 10.1109/CVPRW.2015.7301382.
- [15] G. Cheng, X. Xie, J. Han, L. Guo, and G. -S. Xia, "Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020, doi: 10.1109/JSTARS.2020.3005403.
- [16] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2016, doi: 10.1109/LGRS.2015.2499239.
- [17] S. Thirumaladevi, K. V. Swamy, and M. Sailaja, "Multilayer feature fusion using covariance for remote sensing scene classification," *Acta IMEKO*, vol. 11, no. 1, pp. 1–8, 2022, doi: 10.21014/acta_imeko.v11i1.1228.
- [18] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7894–7906, 2019, doi: 10.1109/TGRS.2019.2917161.
- [19] J. R. Schott, "Principles of multivariate analysis: A user's perspective," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 657–658, 2002, doi: 10.1198/jasa.2002.s479.




- [20] B. Petrovska, E. Zdravevski, P. Lameski, R. Corizzo, I. Stajduhar, and J. Lerga, "Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification," *Sensors*, vol. 20, no. 14, 2020, doi: 10.3390/s20143906.
- [21] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *GIS '10: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010, pp. 270-279, doi: 10.1145/1869790.1869829.
- [22] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 197-209, 2018, doi: 10.1016/j.isprsjprs.2018.01.004.
- [23] B. Zhao, Y. Zhong, G. -S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 2108-2123, 2016, doi: 10.1109/TGRS.2015.2496185.
- [24] Y. Liu, Y. Zhong, F. Fei, Q. Zhu, and Q. Qin, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sensing*, vol. 10, no. 3, 2018, doi: 10.3390/rs10030444.
- [25] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175-2184, 2015, doi: 10.1109/TGRS.2014.2357078.

BIOGRAPHIES OF AUTHORS






Thirumaladevi Satharajupalli    research scholar Jawaharlal Nehru Technological University Kakinada, Kakinada. Asst. Professor in the Department of ECE, KKR & KSR Institute of Technology & Sciences, Guntur. Having 10 years of experience in teaching and Publishing work in various reputed journals national and international. Area of research in digital image processing. She can be contacted at email: thirumaladevice@gmail.com.



Kilari Veera Swamy    Dr.K. Veera Swamy is currently working as Professor in ECE Department, Vasavi College of Engineering, Hyderabad. Earlier he worked as a Professor and Principal at QIS College of Engineering & Technology, Ongole, India. He received his M. Tech and Ph.D. from JNTUH, India. He has twenty-five years of experience in teaching undergraduate students and post graduate students. Two scholars received Ph.D under his guidance. He published 54 research articles (International 51 and National 3) in several reputed journals. He attended 44 international and 11 national conferences. He received 2 patent grants. He published 3 patents. He received Best Teacher Award from JNTUK, Kakinada for the AY:2014-15. He Executed one RPS, one MODROBS, and one Consultancy Project. His research interests are in the areas of image compression, image watermarking, machine learning, antennas, and networking protocols. He can be contacted at email: k.veeraswamy@staff.vce.ac.in.



Maruvada Sailaja    Professor, Department of ECE, Jawaharlal Nehru Technological University Kakinada, Kakinada. Publishing various journals and papers from national and international. Having 20 years of experience in teaching and industry. She is guiding several research scholars her current research interests include areas of research in communications, networks and signal processing. She can be contacted at email: s.maruvada@rediffmail.com.