A Novel Approach for High-Resolution Coastal Areas and Land Use Recognition from Remote Sensing Images based on Multimodal Network-Level Fusion of SRAN3 and Lightweight Four Encoders ViT

Muhammad Kashif Bhatti, Muhammad Attique Khan, *Member IEEE*, Saima Shaheen, Ameer Hamza, Ali Arishi, Dina Abdulaziz AlHammadi, Shabbab Ali Algamdi, Yunyoung Nam, *Member IEEE*

***Abstract*— Land use land cover classification from satellite images (remote sensing) has shown many efforts from the last decade due to ecological surveillance, rapid urbanization, law enforcement, climate change, agriculture drought, and disaster recovery. The low-resolution remote sensing images impact on the accurate prediction; therefore, the high-resolution deep learning architecture is widely required. This article proposes a new deep network-level fusion approach that merges a Stacked Residual Self-Attention CNN (SRAN3) with a lightweight ViT based on 4-encoders to enhance the model performance while reducing computational costs. The SRAN3 model is proposed for extracting sophisticated prominent features, while the 4-encoder-based ViT facilitates effective learning with reduced computation time. These networks are fused using a depth concatenation approach that effectively integrates the strengths of both architectures. The fused model hyperparameters are selected through Bayesian Optimization (BO), significantly improving the learning process. The trained model is later utilized in the testing phase, extracting features from the depth-concatenation layer. The extracted features are fed to neural network classifiers and obtain the final prediction. Two publicly available datasets, EuroSAT and NWPU_RESIS45, are employed to obtain improved testing and validation accuracy. The proposed SRAN3+WNN (Wide Neural Network) and 4-encoder ViT+WNN obtained 96.9% and 92.6% of accuracy; however, the proposed fused network+WNN achieved the highest accuracy of 98.4% on EuroSAT and 94.7% accuracy on the NWPU_RESIS45 dataset, respectively. Also, the proposed fused model interpretation is performed using the explainable artificial technique (XAI), which has shown improved land use and land cover classification.**

***Index Terms*— Remote Sensing; Super Resolution; Residual Self-Attention CNN; SRAN3; Network level Fusion; Customize Vision Transformer; Neural Networks.**

Funding: This study was supported by the grant of the National Research Foundation of Korea funded by the Korean government (MSIP) (No. NRF-2021R1A5A8029876) and the Soonchunhyang University Research Fund. The author, Ali Arishi, at King Khalid University would like to acknowledge the support by the Deanship of Scientific Research through King Khalid University, Saudi Arabia funded by the Large Group Research Project RGP2/267/46. Also, this work was supported through Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R508), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

(Corresponding author*: Yunyoung Nam (ynam@sch.ac.kr), Muhammad Attique Khan (attique.khan@ieee.org).

Muhammad Kashif Bhatti and Saima Shaheen is with Department of CS, HITEC University, Taxila, Pakistan (Kashif.bhatti@hitecuni.edu.pk; saima.shaheen@hitecuni.edu.pk)

Muhammad Attique Khan is with department of AI, Prince Mohammad bin Fahd University, Al-Khobar, KSA (attique.khan@ieee.org)

Ameer Hamza is with Centre of Real Time Computer Systems, Kaunas University of Technology, Lithuania (ameer.hamza@ktu.edu)

Ali Arishi is with Department of Industrial Engineering, King Khalid University, Abha, 61421, Saudi Arabia (awaje@kku.edu.sa)

Dina Abdulaziz AlHammadi is with Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O.Box 84428, Riyadh 11671, Saudi Arabia (daalhammadi@pnu.edu.sa)

Shabbab Ali Algamdi is with Department of Software Engineering, College of Computer Science and Engineering, Prince Sattam bin Abdulaziz University, Al Kharj, Saudi Arabia (s.algamdi@psau.edu.sa)

Yunyoung Nam is with Department of ICR Convergence, Soonchunhyang University, South Korea. (Corresponding author e-mail: ynam@sch.ac.kr).

I. Introduction

Satellite imaging is essential in several fields, such as ecological surveillance, enforcement of laws, coastal areas, and disaster recovery [1, 2]. To complete these tasks, it is necessary to use a manual approach to identify the facilities and products depicted in the images [3]. This technology is critical for companies and government agencies attempting to create precise representations of the Earth, especially for the coastal and land use land cover [4, 5]. The method of accumulating and investigating data about an object, place, or event from a distance is called remote sensing (RS). The technology employed to acquire these images is classified as remote sensing (RS) [6, 7]. Satellite imaging sources often become available at no cost, offering a wide range of coverage and frequent updates with high temporal precision across a vast geographical region [8, 9]. Remote sensing encompasses extensive data collection, processing, and application methods. Additionally, the interpretation of remotely uncovered data frequently requires the application of technology from various disciplines, such as machine learning and pattern recognition [10].

Advancements in remote sensing (RS) technology have facilitated the fast acquisition of an extensive amount of satellite data [11, 12]. The high-resolution satellite images

continuously pose new problems to researchers in remote sensing. Remote sensing (RS) computer vision researchers have extensively employed images for a variety of semantic tasks, including the classification of land cover [13], land use mapping and monitoring [14], geology [15], Deforestation detection, the planning of earth management [16], costal area recognition [17], and the classification of agricultural land [18, 19]. Due to its significant applications in urban planning, Drought Monitoring, and earth hazards, the land cover classification has prompted considerable demand for the field of computer vision [20, 21]. The literature has presented a variety of methodologies for the classification of land cover using remote sensing (RS) images [22-24]. The strategies that have been discussed depend on both supervised and unsupervised learning [25, 26]. Unsupervised learning frequently employs clustering techniques for classification, including fuzzy C-means [27], apriori algorithm [28], and K-means [29]. Still, these methods are unsuitable for managing many labeled imaging RS images.

Conventional methods, including the classification of features by machine learning classifiers and the manual construction of features, were employed in the supervised learning techniques. The handmade features are obtained using prior knowledge, including details about texture, shape, and point features [30]. Computer vision significantly influences the development of an intelligent system [31]. The three primary components of computer vision are feature extraction, fusion, and selection. Therefore, feature fusion techniques were introduced by computer vision researchers. The fusion procedure has improved the accuracy of predictors in the system. Parallel and serial-based fusion [32] are the most frequently employed fusion techniques. The primary constraint of this phase is that it increases in computation time. In the fusion process, researchers typically combine the features of two models. Still, this method is time-consuming and inefficient. Several researchers employed feature selection methods to eliminate extra features from the final classification and resolve this issue.

Although it is challenging to identify the most offensive features from complicated scenarios, deep learning techniques are more reliable and show better performance for complex classification tasks. Deep learning has seen significant worldwide success in a wide range of applications, including object classification and remote sensing (RS) [33]. Deep learning stands out due to its remarkable ability to learn and perform well on large datasets. One well-known deep learning model that can extract abstract information from images in a hierarchical manner is the convolutional neural network (CNN) [34]They have been used in several industries, such as face detection, irrigation management, satellite imaging, and healthcare. A substantial amount of labeled training data is needed to train CNN and obtain satisfactory performance effectively.

Several techniques have been introduced in the literature for land use land cover classification from remote sensing images. They focused on computer vision [35], especially on deep learning techniques for classification. These techniques can be employed for various remote sensing applications, including historical buildings and identifying land use areas [36]. Pritt et al. [37] presented a deep learning-based approach for classifying satellite images. Using high-resolution, multispectral satellite images, they carried out object and facility recognition using the methods they described. With this method, they were able to achieve 95% accuracy. The system's drawback was the state-of-the-art object detection technology, which performs poorly for satellite images. Gao et al. [38] presented a region-based deep learning technique that was suggested for segmenting satellite images. In this method, they used rooftop detection by employing the segmentation strategy. From this strategy, they obtained 92% accuracy. This provided method could not avoid the speckle-like inaccuracy sometimes encountered in the segmentation model. Audebert et al. [39] suggested a technique that uses CNN architecture to combine data from heterogeneous sensors (optical and laser) and used various kernels to improve the accuracy of semantic segmentation classification. The image segments in the same investigation were created using multiresolution segmentation. They used images from Semarang, Indonesia, the NDVI, homogeneity, brightness, and rectangle fit features for each segment, which were acquired and used as deep learning input. Nguyen et al. [40] presented a deep-learning technique for classifying satellite images. They use deep neural network architecture based on CNN and attained an accuracy of 93%. The method's drawback is that processing satellite images to map the agricultural framework is complex. Zhang et al. [41] presented building height extraction from satellite images. They presented a method for forecasting bottom elevation combined with a DSM-based approach and stereo-matching technique. Hasan et al. [42] presented a new resource allocation method in 5G heterogeneous networks. The authors developed a novel dynamic subcarrier allocation algorithm based on biogeography to minimize crosstire subcarrier snooping issues in MeNB and HeND. They were able to attain 83.6% spectral efficiency and 88.1% outage. It was beyond the capabilities of the methods in use. Chen et al. [43] designed a deep learning-based method for identifying remote sensing data. They used attention-guided sparse filters to build a network architecture based on CNN. Consequently, they had a 94% accuracy rate. The technique's drawback was the absence of a sizable dataset. In Alam et al. [44] presented proximal sensor for the classification of land use in coastal areas. They employed SVM for the whole experiments and they achieved 84% accuracy. The limitation is was the noisy dataset. Albarakati et al. [45] presented a unified super resolution based technique for the classification of remote sensing images. The authors used RSI-CB128 and WHU dataset for the experimental process and they achieved 95.7%, 97.5% of accuracies. For the hyperparameters tuning, the authors was employed Bayesian optimization. This method was the limitation of proposed method due to its high time for training.

Deep learning techniques for classifying remote sensing data. However, they are still facing an issue of accurate land use and land cover prediction from remote sensing images due to high similarity among several classes, as shown in Figure 1. The major problem of these models is extracting important features from a single CNN model; therefore, examining a more efficient approach, such as network-level fusion of CNN architecture instead of features-level fusion. Two models are generated and integrated into the network-level fusion process

to enhance accuracy and reduce computational costs. However, the total learnables are increased or somehow added in the network-level fusion, which is also an issue for model training using fewer resources. We proposed a deep network level fusion and residual attention mechanism to consider these challenges and accurately predict LULC from remote sensing images. The proposed network-level fusion process involves creating two custom networks based on the residual attention learning and customized vision transformer. These networks are subsequently concatenated using a depth concatenation layer.

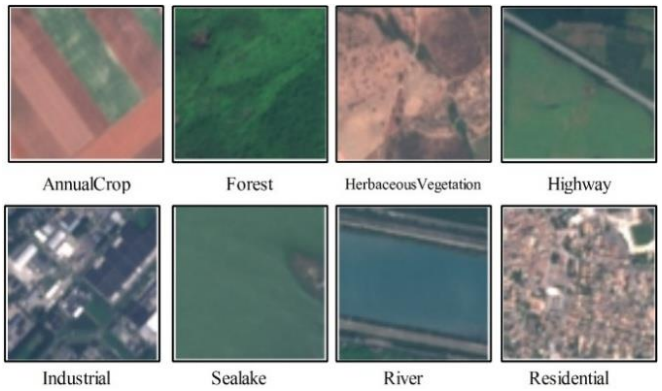


Figure 1: Samples of remote sensing images collected from the EuroSAT dataset [46]

We proposed a deep network level fusion and residual attention mechanism to consider these challenges and accurately predict LULC from remote sensing images. The proposed network-level fusion process involves creating two custom networks based on the residual attention learning and customized vision transformer. These networks are subsequently concatenated using a depth concatenation layer. The primary contributions of this work are as follows.

- We proposed a novel Stacked Residual Self-Attention CNN (SRAN3) Network for the refined prominent information extraction.
- We proposed a lightweight vision transformer based on 4-encoders (LViT-4E) for better learning and efficient computation time.
- We proposed network-level fusion based on SRAN3 and 4-encoder-based ViT to merge the capabilities of both models.
- An ablation study is conducted along with the performance of proposed models on various configurations and compared with the state-of-the-art models.

II. PROPOSED WORK

The proposed deep learning architecture for LULC classification from remote sensing images has been presented in this section. The proposed architecture started with dataset augmentation through statistical techniques. After that, two novel models are proposed, such as SRAN3 and 4-encoder-based ViT. Both models are fused using a depth-concatenation layer and trained on the augmented dataset (training set) with optimized hyperparameters. The next step is testing the trained model using a testing image set and obtaining final prediction results such as recall, precision, and accuracy. Each listed step is presented in Figure 2 and further discussed in the below sub-sections.

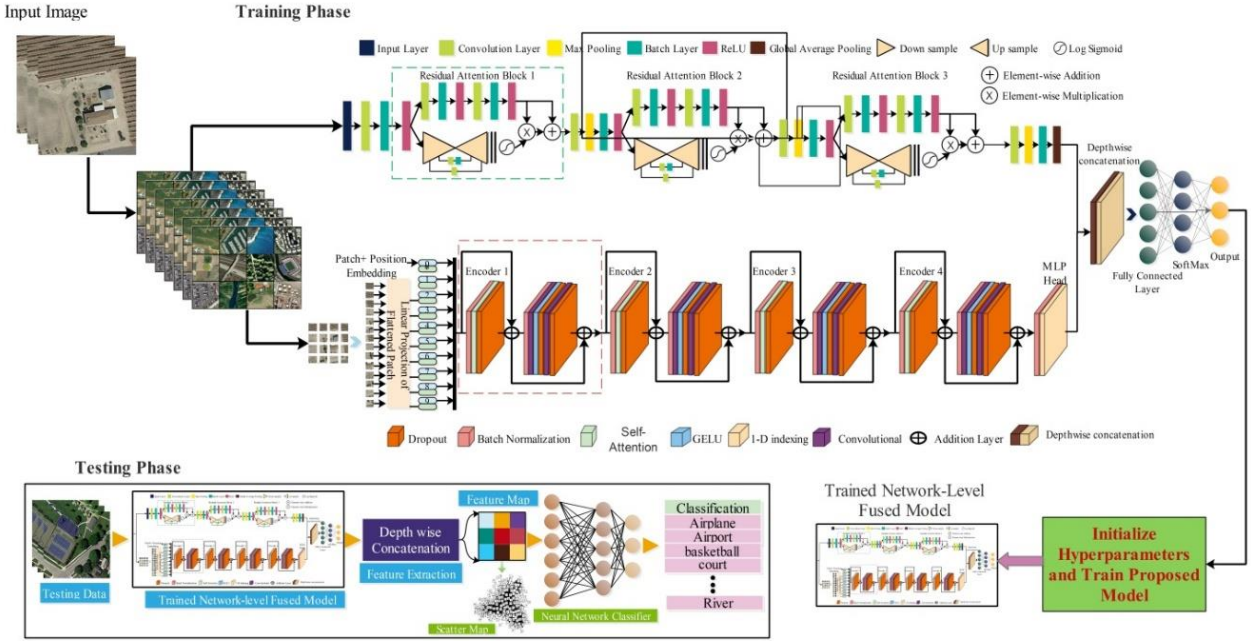


Figure 2: Proposed deep learning architecture for High-Resolution Satellite Land Use classification based on SRAN3 and 4-Encoder ViT integration

A. Dataset

This work utilizes the publicly available dataset for experimental work. The selected datasets are Eurostat (https://www.kaggle.com/datasets/apollo2506/eurosat-dataset/) and NWPU_RESISC45

(https://figshare.com/articles/dataset/NWPURESISC45_Dataset_with_12_classes/16674166).

Eurostat: The Eurostat dataset consists of ten classes: Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, and Sea Lake. The data of these classes was collected from the satellite with a dimension of $64 \times 62 \times 3$ and a distance from the ground of 10m. The size of a single image is $256 \times 256 \times 3$ with $96 \times 96 \times 3$ dpi. The total count of images in the dataset is 27000. A few sample images of this dataset are shown in Figure 1.

NWPU_RESISC45: The NWPU_RESISC45 dataset consists of 10500 images with 12 classes: airfield, Anchorage, beach, dense residential, farm, Flyover, forest, game space, parking space, river, sparse residential, and storage Cisterns. The resolution of images in this dataset is$256 \times 256 \times 3$ and it has 24 dpi. A few sample images of this dataset are shown in Figure 3.

Table 1 presents a detailed summary of these selected datasets. In this table, original and augmented images are presented. Also, 70% of the images are used for training, and 30% are used for testing. Before data augmentation, both datasets are divided into a 70:30, and then augmentation has been performed for the training set. For the EuroSAT dataset, the testing images are also augmented. In contrast, for the NWPU dataset, only a few classes are considered for augmentation, such as Anchorage, Forest, River, and a few more.

Figure 3: Samples images of the NWPU remote sensing dataset

Table 1: Description of EuroSAT and NWPHU_RESISC45 remote sensing datasets

EuroSAT Dataset			**NWPU_RESISC45**		
Classes	**Original Count**	**Augmented Train/Test**	**Classes**	**Original Count**	**Augmented Train/Test**
Annual Crop	3000	7000/3000	Airfield	1400	1634/700
Forest	3000	7000/3000	Anchorage	700	1634/700
Highway	2500	5834/2500	Beach	700	1636/698
Industrial	2500	5834/2500	Dense residential	700	1636/698
Pasture	2000	7000/3000	Farm	1400	1634/700
Permanent Crop	2500	5834/2500	FlyOver	700	1984/350
Residential	3000	7000/3000	Forest	700	1634/700
River	2500	5834/2500	Game space	1400	1635/699
Sea Lake	3000	7000/3000	Parking space	700	1636/698
Herbaceous Vegetation	3000	7000/3000	River	700	1634/700
Total EuroStat	27000	65336/27000	Sparse residential	700	1636/698
Total NWPU-45	10500	17984/8039	Storage Cisterns	700	1636/698

B. Proposed HR-SAN²

The second order Attention Network (SAN) model is an advanced deep learning architecture for high-performance image super resolution tasks. By using second-order functional statistics, it capture complex spatial dependencies and improve the representation of functionalities. The SAN model is particularly effective in situations requiring the restoration of fine texture and noise suppression. In this work, we proposed a 2 block based lightweight SAN model to improve the resolution of remote sensing images. Remote sensing images are often confronted with challenges such as low spatial resolution, atmospheric disturbances, noise, and lighting changes. It is important to improve these images for tasks such as terrestrial cover classification, disaster monitoring, and satellite image detection. The proposed HR-SAN² architecture efficiently captures complex spatial dependencies and improves performance learning using second-order statistics to make it stronger in dealing with inherent challenges in remote sensing data. The proposed HR-SAN² accepts the input size of $224 \times 224 \times 3$. The image pass to the initial convolutional layer which configured with 3×3 kernal size, 1×1 stride, 32 channels. The convolutional operation is $f_{map} = \varphi(X_{LR})$, where the φ is convolutional operation, $X_{I,R}$ is the low resolution input image. After that, first SAN block is employed. In this block, a convolutional layer is followed by the second order channel attention module and residual activation. The mathematical modeling of SOCA module is:

$$f_{map} = SOCA(\varphi\left(f_{map}\right) + X_{LR} \quad (1)$$

$$S_o = \frac{1}{HW} f_{flat} f_{flat}^T \quad (2)$$

$$C_\tau = \sigma(FC(\mu((S_o)))) \quad (3)$$

Where S_o is the SOCA module operation, f_{flat} belong to $\mathbb{R}^{D\times(H\times W)}$ is used for reshaped the feature map. The covariance matrix S_o is reduced and attention weights C_τ are measured by performed sigmoid activation and dense layer on S_o. The weights are further refined using multiplication between the feature map and attention weights which is indicated with f_{ref}.

$$f_{ref} = f_{map}.C_\tau \quad (4)$$

After this, second block is employed with the same mechanism. A addition layer is employed to add the initial convolutional and output feature map of second SOCA module, which is formulated the skip connection to ensure the stable training and feature preservation. We have;

$$\psi_{add} = f_{map} + f_{ref} \quad (5)$$

Where ψ_{add} presents the addition layer. In the end a convolutional layer is performed to mapped based the refined features to the high resolution space, which is defined as:

$$X_{HR} = \varphi(\psi_{add}) \quad (6)$$

the proposed have 12 layer with 1.3 million parameters. The outcomes of the proposed architecture is shown in Figure 4.

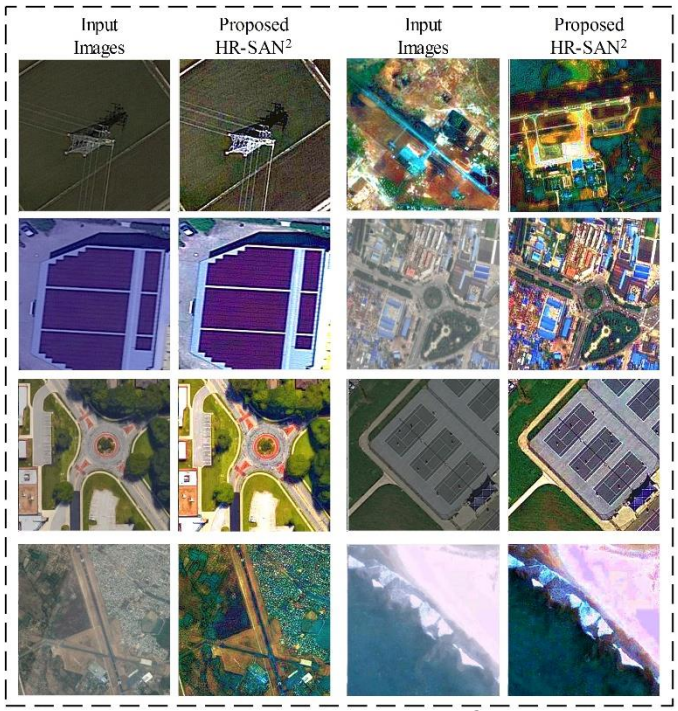


Figure 4: Samples of proposed HR-SAN^2 module for high resolution remote sensing images

C. *Proposed Stacked Residual Self-Attention CNN (SRAN3)*

Convolutional Neural Networks (CNNs) are the kind of artificial neural networks that are designed to interpret spatial input, specifically images. In the CNN, a convolutional layer is employed to extract important information and investigate input data. We proposed stacked residual self-attention (SRAN3) CNN architecture in this work. The detailed architecture is shown in Figure 5. In this figure, it is noted that three residual attention blocks have been added. Attention modules are typically essential to strengthening Convolutional Neural Networks' learning capabilities by focusing their attention on useful information and discarding irrelevant background elements. This relevant information pertains to the target objects or specific class areas within an image that require classification and localization in the context of object detection.

The attention mechanism architecture comprises critical components, including a convolutional operation, a Multi-Layer Perceptron, and a sigmoid activation that generates a mask determined by the input feature map. The proposed SRAN3 starts with an image input size of $227 \times 227 \times 3$. Subsequently, a convolutional layer with 3×3 filter size, 2×2 stride, and 32 channels and batch normalization are performed. After that, three residual self-attention modules are stacked using some transitional layers. The transitional layers are convolutional, max pooling, batch normalization, and ReLU. The Residual self-attention is established by stacking multiple attention modules. Each is separated into two portions: the core and mask branches. The structuring of attention modules facilitates the learning process through their efficient layout. This approach preserves the beneficial feature of the low features while allowing the possibility to bypass the mask branch and directly proceed to higher layers. The multilayered associative memories (AMs) structure enables an incremental strengthening of feature mappings. After these modules, one convolutional is employed and configured with a 1×1 filter size, 1×1 stride, and 512 channels. A global average pool, fully connected, softmax, and classification layers are employed at the network's end. The loss function of the proposed SRAN3 is cross entropy. The mathematical formulation of the loss function is defined as:

$$SRAN_{Loss} = \frac{1}{B}\sum_{x=1}^{B}\sum_{y=1}^{T} g_{xy}\log(q_{xy}) \quad (7)$$

Where B presented the observation in the batch, T is the number of target classes and $g_{xy}$is denoted by the correct classification x for instance$y$. q_{xy} Is the predicted probability. The proposed network SRAN3 has 86 layers in total and 7.1M trainable parameters.

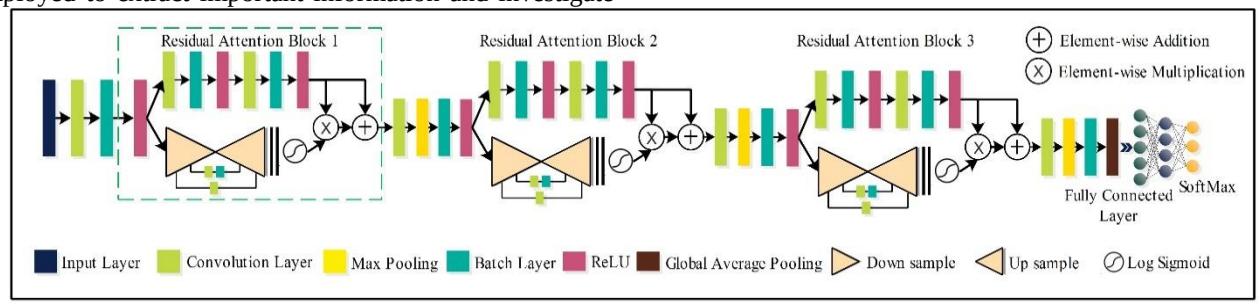


Figure 5: Proposed architecture of SRAN3 for LULC classification from RS images.

1) *Residual Attention Module*

The residual attention module generates a weight matrix to analyze the connections among various channels within a

feature map and reduce its duplication. Therefore, the input feature map $f_\alpha \in I^{H\times W\times C}$, Two convolutional layers are performed, and the spatial dimension remains the same, but one of the feature maps covers half of the channels.$\{Conv(f_\alpha), Conv(f_\alpha)\} \in \left\{ I^{H\times W\times C}, I^{H\times W\times \frac{C}{2}} \right\}$, after that, we applied a transpose convolutional layer to hold back a quarter of the spatial size. Mathematically, the transpose convolutional is defined by Equation (8):

$$C[a,b] = \sum_{i=0}^{H_k-1} \sum_{j=0}^{W_{k-1}} \tau[a'+i, b'+j].\rho[i,j] \quad (8)$$

Where $[a', b']$ is presented as the input position, a' and b'adjusted corresponding to the stride, $\rho[i,j]$ denoted the kernel value at the position (i,j), H_k and W_k are the height and width of the kernel, and $C[a,b]$ is the output of this equation, respectively.

After transposing, a reshape operation is performed, and softmax activation is employed as:

$$f_{softmax_{i,j}} = \frac{e\left(T_{conv}(f_\alpha)_i^1 \, (T_{conv}(f_\alpha)_j^2)^R\right)}{\sum_{i=1}^{\frac{C}{2}} e\left(T_{conv}(f_\alpha)_i^1 \, (T_{conv}(f_\alpha)_j^2)^R\right)} \quad (9)$$

Where $f_{softmax_{i,j}} \in I^{C\times\frac{C}{2}}$ is denoted the i^{th} channels impact on j^{th} channel. The Softmax operation is employed row-wise. After this, the element-wise multiplication uses a residual connection with the input map. Mathematically, it is formulated as follows:

$$f_c = f_\alpha \oplus (f_{softmax} \otimes Conv(f_\alpha)) \quad (10)$$

Where $\oplus$ presented is the element-wise addition, $\otimes$ denoted the element-wise multiplication, and f_c is the final output of the residual attention module regarding channels. Figure 6 presents its internal architecture.

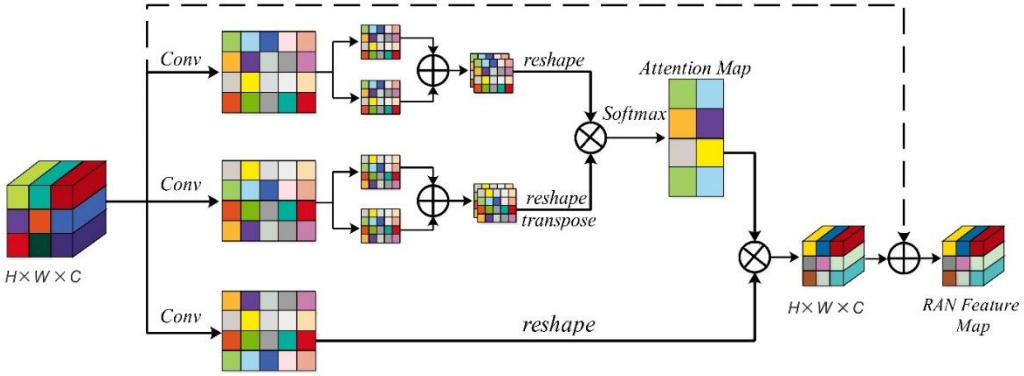


Figure 6: Internal architecture of residual attention module

D. *Customized 4Encoder-based ViT*

A Vision Transformer (ViT) is a deep learning model that employs transformer architecture, typically used in natural language processing (NLP). This is dissimilar to traditional convolutional neural networks (CNN) that analyze images using convolutional layers. CNN is very effective in detecting and analyzing small-scale patterns in images; however, they have challenges when it comes to understanding long-distance relationships and overall context because they primarily focus on local areas of the image. Additionally, they impose a rigid hierarchical framework that limits flexibility to learn features and includes significant inductive biases that may not effectively apply to various patterns. Therefore, ViT resolves these concerns using the transformer architecture, which splits an image into patches and treats them as tokens. This method utilizes self-attention processes to capture associations across the whole image, allowing for improved modeling of the overall context and distant relationships while minimizing the need for inductive biases. There are various pre-trained vision transformers for the classification problem, such as base 16, small 16, large 16, and many more. The pre-trained large 16 ViT consists of 24 transformers with 86.8 parameters.

In this work, we designed another customized lightweight vision transformer (LViT) consisting of 4 transformers with 25.8 Million parameters. The proposed model starts with an input layer that consists of $224 \times 224 \times 3$. A patch embedding layer is added with a 16×16 patch size. After that, a position embedding layer is added to maintain the spatial structure of the image. Subsequently, an addition layer is employed to merge the information of both patches and their position concerning the input image. After that, the first transformer block is started based on the dropout layer with 0.1, batch normalization, multi-head self-attention layer, and dropout layer 0.1. The information from previous layers is added to these layers by employing the addition layer. In the next batch normalization, two convolutional layers, one GeLU activation, and two dropout layers are added. The following three transformers are based on the same phenomena as the first transformer. After the fourth transformer 1D indexing, fully connected, softmax, and classification layer is added. Mathematically, it is transformed as follows:

Input tokens (u) are linearly mapped as queries (Q), key (K), and value matrix (V). The weights coefficient matrix is computed for Q and K matrix and use the obtained matrix to compute the weighted sum of vectors in K that is a weighted integration between the tokens. Based on this, the self-attention (SA) is computed through the following equation:

$$AT(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \quad (11)$$

Where dimensions of the key vector are denoted by $\sqrt{d_K}$ and T is a transpose of QK matrix. After that, the multi-head self-

attention is performed to concatenate the h-heads by the following equation:

$$H_i = AT\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (12)$$

$$MHSA(Q,K,V) = Concatenate(H_1,\ H_2, \dots, H_h)W^o \quad (13)$$

Where W_i^Q, W_i^K, W_i^V, and W^o denotes projections that are trainable parameter matrices and H denotes the number of $MHSA$ heads. Combining these all, a complete transformer block is given as:

$$t(u) = MHSA(NorL(u)) + u \quad (14)$$

$$Transform(u) = MLP\left(NorL(t(u))\right) + t(u) \quad (15)$$

Where u denotes the input tokens and $NorL$ is a normalization layer that is applied before $MHSA$ and MLP. The MLP contains two FC layers and residual links, along with GELU activation functions. The designed model has 54 layers and four encoders. The architecture of the proposed ViT is visually presented in Figure 7.

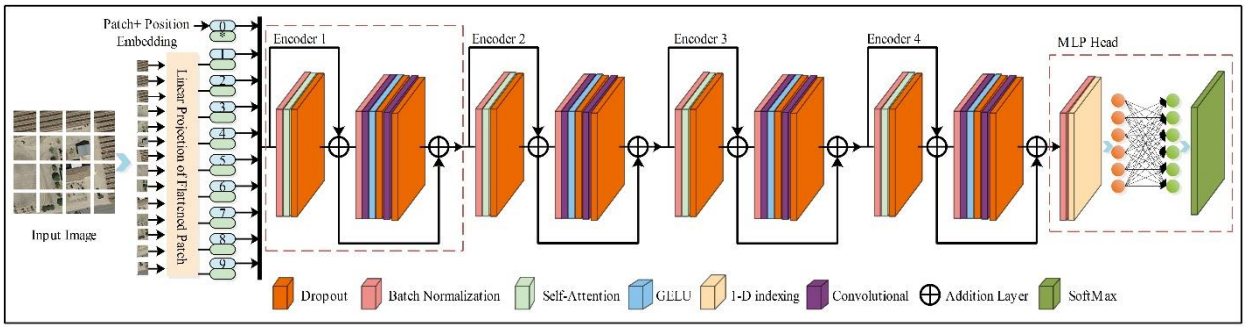


Figure 7: Architecture of proposed 4 encoder-based lightweight vision transformer (LViT)

E. Network Level Fusion and Training

The integration of multiple neural networks to leverage their combined resources is referred to as network-level fusion in deep learning [47]. By using network-level fusion, the model may acquire information from a wider range of features. This fusion technique may assist in locating complex patterns and correlations in data that may go unnoticed by a single network. By incorporating features and decision-making abilities, this technique has the potential to significantly improve both stability and accuracy.

In this work, we proposed network-level fusion based on proposed networks (SRAN3 and 4Encoder-based ViT). The architecture of both models is fused using the depth concatenation layer, as shown in Figure 8. It merges the different tensors from the different networks. Let's suppose we have two feature maps φ_{F1} with the shape of $[B_N, H, W, D_1]$. φ_{F2}With shape of $[B_N, H, W, D_2]$. B_N presented as the batch number of samples. H indicates the height, W is the width and D is presented as the depth $D \in \{D_1, D_2\}$. The concatenation operation is defined by Equation (16).

$$Y_c = \begin{Bmatrix} \varphi_{F1} & for\ \theta < D \\ \varphi_{F2} & for\ \theta \geq D \end{Bmatrix} \quad (16)$$

Where θ is the depth index in the concatenated map feature map Y_c. The resultant shape of the feature is

$$\omega = [B_N, H, W, D_1 + D_2] \quad (17)$$

In our scenario, the dimension of the extracted feature of the proposed SRAN3 is $1 \times 1 \times 512$, and the 4Encoder-based ViT has $1 \times 1 \times 512$. After employing the depth concatenation layer, the obtained feature map is$1 \times 1 \times 1024$. After the complete design of this fused network, training was performed using augmented datasets, as discussed in Section 2.1 and Table 1. The hyperparameters for the proposed model training were the learning rate, mini-batch size, momentum, learning rate decay, epochs, loss function, and optimizer having values of 0.00143, 32, 0.0912, 0.01, 100, categorical cross-entropy, and SGDM. These hyperparameters, such as initial learning rate, momentum, and learning rate decay, are selected through Bayesian Optimization (BO) [48]. After that, the proposed model is trained and later utilized in the testing phase for the final classification performance and visual prediction.

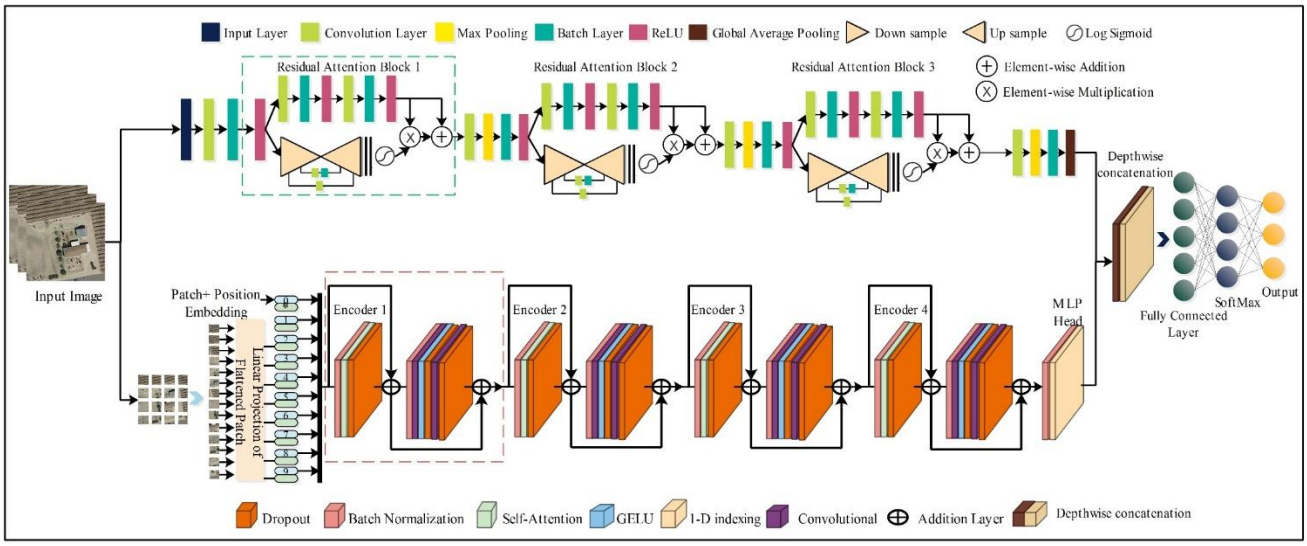


Figure 8: Architecture of proposed network-level fusion based on proposed models (SRAN3 and 4Encoder-based ViT)

F. Feature Extraction and Testing

Convolutional Neural Networks often return feature maps classified into three tiers: low-level, mid-level, and high-level. Each of these layers has different types of information. Low-level feature maps include fundamental patterns such as textures, edges, and corners. High-level features provide exclusive and conceptually useful information, while mid-level feature maps include abstract and more ordered patterns, such as specific sections of an object or texture. Because of their low dimensions, CNNs' deep layer produces high-level features. The low dimension reduces computing complexity and memory requirements.

Our work employs the test data for the feature extraction process after obtaining a trained fused model. The average pooling activation is utilized from the SRAN3 models, and the extracted feature size is $N \times 512$. From the proposed 4Encoder-based ViT, the 1D indexing layer is employed, and the dimension of features is $N \times 512$. Moreover, both models are network-level fused, and fused features are extracted from the depth-wise concatenation node. The obtained feature size is $N \times 1024$, which was further passed to neural network classifiers for classification accuracy. The entire testing process is illustrated in Figure 9. In this figure, it is noted that the testing query images are passed to the trained network and extracted deep features from the depth concatenation layer that are further passed to neural network classifiers and, as a result, returned a specific label such as Airplane, Airport, and court.

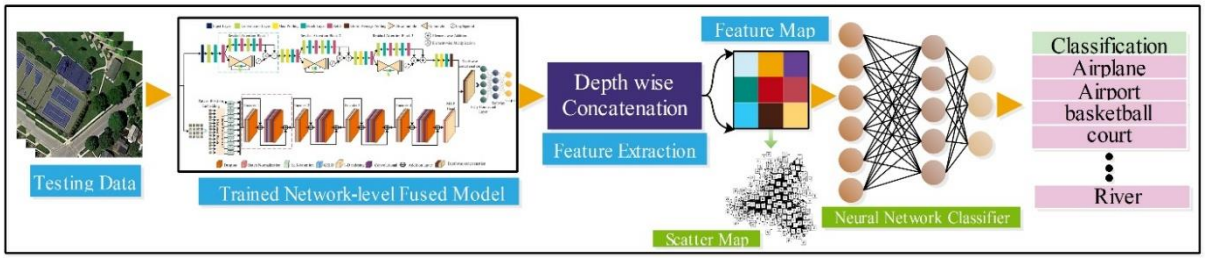


Figure 9: Feature extraction and evaluation process using the proposed models

III. Results and Analysis

The experimental setup and the classification results have been discussed in this section. The datasets of this work are divided into three parts: training, testing, and validation. The 60% of the data is employed for the training process, and the 10% of the data is utilized for the validation during the training process. The remaining 30% of the data is used for the testing process. The hyperparameters for the proposed model training were the learning rate, mini-batch size, momentum, learning rate decay, epochs, loss function, and optimizer having values of 0.00143, 32, 0.0912, 0.01, 100, categorical cross-entropy, and SGDM. The proposed models are trained using the 10-fold cross-validation to improve the generalization. For the classification phase, neural network classifiers with a different number of hidden layers are employed. The LOOCV method is employed to evaluate neural network classifiers. The performance of neural network classifiers is evaluated using precision, recall, f1-score, accuracy, and prediction time. All the experiments are simulated on a Desktop system configured with core i5 with 3.2GHz, 64 GB RAM, and NVIDIA RTX 3060 12 GB.

A. Results of EuroSAT Dataset

Experiment 1: In the first experiment, the 4Encoder-based ViT model is designed from scratch and trained on the

EuroSAT dataset. The trained model is employed for the feature extraction and fed to the neural network (NN) classifiers. The results of this experiment are described in Table 2. This table shows that the wide neural network (WNN) classifier outperformed all the listed classifiers. The WNN classifier achieved the highest accuracy of 96.9%, a precision rate of 95.14%, and 94.26% F1-Score, respectively. Figure 10 shows the confusion matrix for further verification of these measures. The numerical analysis was also conducted for the other classifiers, and it was noted that the medium neural network (MNN) classifier also obtained a better accuracy of 95.3%. The computation time is indicated on the proposed 4encoder features, and it is perceived that the narrow neural network (NNN) takes less time, which is 712.75 (sec), while the MNN requires 936.0 (sec), which is the highest one.

Experiment 2: In the next experiment, the proposed SRAN3 CNN architecture is tested separately, and features from the global average pooling layer are extracted. The extracted features are passed to the neural network classifiers. The classification results of SRAN3 on the EuroSAT dataset have been presented in Table 2. From this table, it is seen that the WNN classifier achieved the highest accuracy of 92.6%. The other parameters, such as precision rate, are 92.44%, recall rate is 91.63%, and F1-Score is 92.033%, respectively. Moreover, the confusion matrix of this classifier is shown in Figure 11. The computation time is measured for this experiment, and it is noted that the MNN classifier has the shortest time, which is 64.456 (sec), and the TNN classifier has the longest time, which is 291.98 (sec).

Experiment 3: In the final experiment, the designed models of experiment 1 (4Encoder-based ViT) and experiment 2 (SRAN3) are network-wise fused and trained on the EuroSAT dataset for final classification. After training, the trained model is employed for the feature extraction. The extracted information is passed to the shallow neural network classifier. The classification results are described in Table 2. From this table, it is perceived that the WNN classifier achieved an accuracy of 98.4%. The precision rate is 97.21%, the recall rate is 98.40%, and F1-Score is 97.79%, respectively. In Figure 12, a confusion matrix is presented for further authentication. The computation time is also recorded for this experiment, and it is observed that the network-level fusion process significantly decreases the computation of the classifier. The TNN classifier has the highest time, 438.72% (sec), while the lowest is 237.04 (sec).

Compared with experiments 1 and 2, the fused network significantly increased performance by ~1.5%. This method's computation time is slightly higher than the SRANB3 and lower than the proposed 4Encoder-based ViT model. In addition, the precision rate, which is the strength of a fusion process at the network level, is increased.

Table 2: Classification results of the proposed framework on the Eurostat dataset. The bold values presented the highest-noted performance metrics

Classifiers	Proposed ViT	SRAN3	Fused Network	Precision	Recall	F1-score	Accuracy	Time (sec)
NNN	✓	-	-	88.53	88.72	88.60	89.8	712.75
	-	✓	-	89.62	89.41	89.51	89.7	169.96
	-	-	✓	89.14	88.21	88.67	90.2	344.1
MNN	✓	-	-	94.97	95.14	95.00	95.3	936.1
	-	✓	-	91.67	91.43	91.54	91.4	64.456
	-	-	✓	96.54	97.16	96.34	97.8	237.04
WNN	✓	-	-	**95.14**	**94.26**	**95.15**	**96.9**	736.3
	-	✓	-	**92.44**	**91.63**	**92.033**	**92.6**	75.129
	-	-	✓	**97.21**	**98.40**	**97.79**	**98.4**	297.56
BNN	✓	-	-	87.54	87.72	87.60	87.8	898.02
	-	✓	-	88.94	86.14	87.51	89.6	207.47
	-	-	✓	89.96	90.10	90.02	90.5	397.64
TNN	✓	-	-	87.54	87.72	87.60	87.8	874.2
	-	✓	-	89.41	89.13	89.26	89.4	291.98
	-	-	✓	89.11	89.32	89.21	89.7	438.72

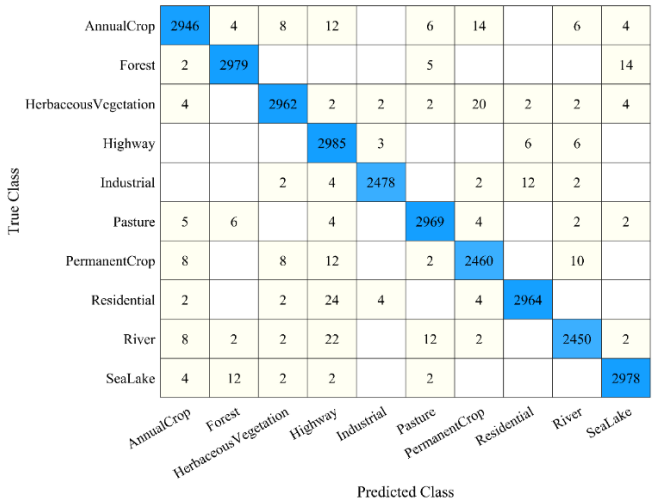


Figure 10: Confusion matrix of MNN classifier using 4Encoder-based ViT CNN architecture that tested on EuroSAT dataset

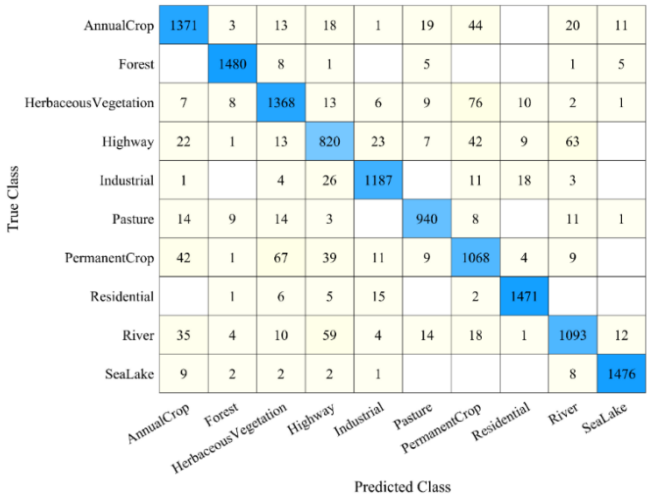


Figure 11: Confusion matrix of proposed RSAN3 architecture using MNN classifier on EuroSAT dataset

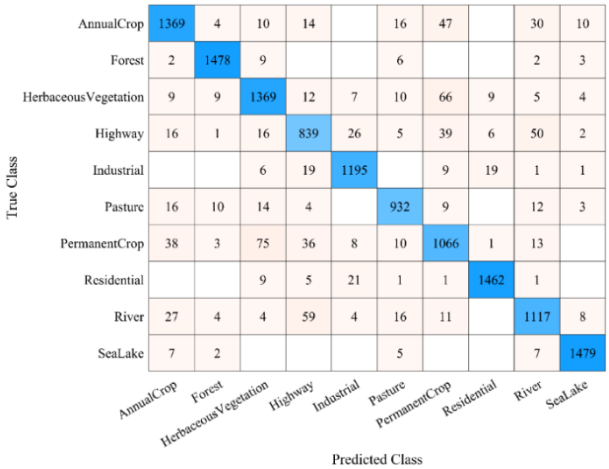


Figure 12: Confusion matrix of proposed fused network model on EuroSAT dataset

B. Results of NWPU_RESISC45 Dataset

Experiment 1: Similar experiments are performed for this dataset, as discussed in Section 3.1. In this experiment, the 4Encoder-based ViT model is trained on the NWPU_RESISC45 dataset and later utilized for feature extraction and classification in the testing phase. The classification results have been presented in Table 3. From this table, it is observed that the WNN classifier obtained 91.3% accuracy, which is higher than the other listed classifiers in this table. The precision rate, recall, and F1-Score are also measured; the values are 90.54%, 90.94%, and 90.73%, respectively. The MNN classifier achieved 90.6% accuracy, ~ 0.7% less than the WNN classifier. Figure 13 shows the confusion matrix for further confirmation of the numerical statistics of the WNN classifier. These parameters are also measured for all the listed classifiers in this table. The computation time of the neural network classifiers is noted in this experiment. The MNN classifier has the lowest execution time, which is 89.118 (sec), while 175.74 (sec) is the highest time for the TNN classifier.

Experiment 2: In this experiment, the designed SRAN3 model is employed for the training on the NWPU_RESISC45 dataset, and the prominent information is extracted in the form of features. The neural network classifiers are utilized to classify extracted features, and results are presented in Table 3. The tabular data shows that the WNN classifier again outperformed and achieved 94.8% accuracy. There are some other computed measures, such as a precision rate of 94.51%, a recall rate of 93.94%, and an F1-Score of 94.22%, respectively. The confusion matrix is also presented in Figure 14. All the numerical statistics are also measured for the listed classifiers. The MNN classifier gained 93.1% accuracy, ~1.7% lower than the WNN. The lowest computation is noted from the WNN, which is 48.422 (sec), and the TNN classifier takes the longest time for execution, which is 158.99 (sec).

Experiment 3: In this experiment, the fused network is used for training on the same dataset, and features are extracted and passed to the neural network classifier as the previous experiments have been done. Table 3 shows the classification results of the fused network. The WNN classifier gained the highest accuracy of 95.7%, whereas the precision, recall, and F1-Score values were 94.41, 94.94, and 95.67%, respectively. The confusion matrix of the WNN classifier is displayed in Figure 15. The computation time is also recorded for this experiment, and it is observed that the MNN classifier is significantly faster than the previous experiments (1 and 2) of this dataset and completed the task in 36.604 (sec). In contrast, the TNN classifier has the longest time, 269.29 (sec). In summary, it is noted that in experiments 1 to 3, the fused network has significantly outperformed the other networks. The fused network is improved by ~0.9 in terms of accuracy, and ~12.0 (sec) is faster in computation time.

Table 3: Classification results of the proposed architecture on the NWPU_RESISC45 dataset

Classifiers	Proposed ViT	SRAN3	Fused Network	Precision	Recall	F1-score	Accuracy	Time (sec)
NNN	✓	-	-	87.46	88.81	88.12	88.3	172.6
	-	✓	-	90.60	90.96	90.27	90.2	85.998
	-	-	✓	91.24	92.46	91.84	92.7	169.04
MNN	✓	-	-	90.14	90.46	90.29	90.6	89.118
	-	✓	-	92.91	93.17	92.49	93.1	84.82
	-	-	✓	93.35	93.14	93.24	94.2	36.604
WNN	✓	-	-	90.54	90.94	90.73	**91.3**	104.01
	-	✓	-	94.51	93.94	94.22	**94.8**	48.422
	-	-	✓	94.41	94.94	95.67	**95.7**	119.17
BNN	✓	-	-	87.97	88.64	88.30	88.3	171.05
	-	✓	-	92.34	91.57	91.95	92.6	93.815
	-	-	✓	92.12	93.64	93.37	93.6	172.1
TNN	✓	-	-	87.31	86.64	86.97	87.58	175.74
	-	✓	-	92.15	91.41	91.77	92.1	158.99
	-	-	✓	91.14	92.24	91.68	92.4	269.29

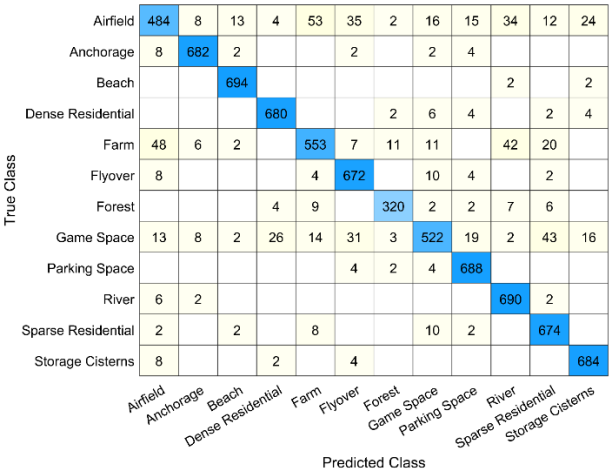


Figure 13: Testing confusion matrix of WNN classifier using proposed 4Encoder-based ViT architecture on NWPU_RESISC45 dataset

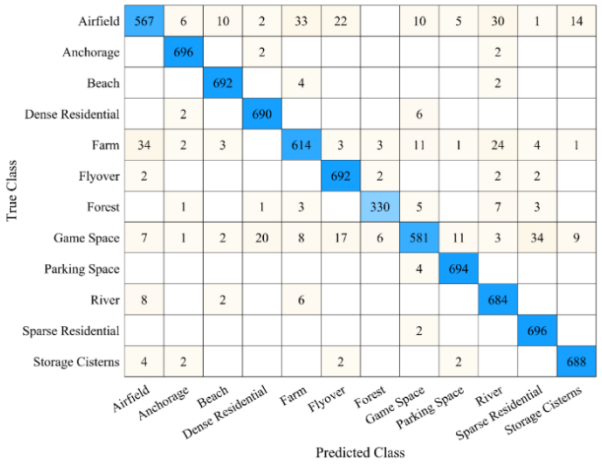


Figure 14: Testing confusion matrix of proposed RSAN3 with WNN classifier on NWPU_RESISC45 dataset

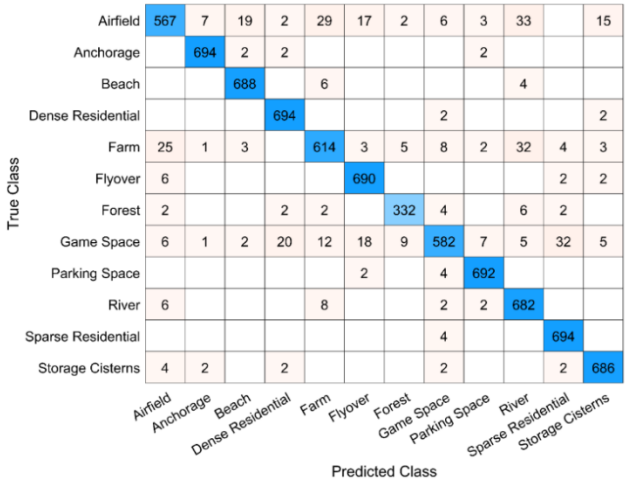


Figure 15: Testing confusion matrix of WNN classifier using the proposed fused network on NWPU_RESISC45 dataset

C. Discussion and comparison

This section presents a detailed discussion of the proposed architecture, which includes prediction time-based analysis, ablation studies, confidence interval-based analysis, model interpretability, and comparison with state-of-the-art (SOTA) techniques.

i. 1) Prediction Time-based Analysis

A brief comparison of the prediction time of each neural network on a EuroSAT and NWPU_RESISC45 dataset has been presented in Figure 16 and Figure 17. The NNN, MNN, and WNN are shallow neural networks with one hidden layer with 10, 25, and 100 neurons, respectively. The BNN neural network has two hidden layers, each with ten neurons. TNN contains three hidden layers, each of which has ten neurons. The prediction time of each neural network classifier is noted and compared to all the neural network classifiers on the EuroSAT dataset, as shown in Figure 16. From this figure, it is projected that the TNN classifier takes the highest prediction time on overall observation, which is ~7000 (obs/sec) on fused network features. The WNN takes the lowest prediction time for all observations, ~3900 (obs/sec), using SRAN3 network

weights. The remaining classifiers inference time are lies among the ~4000 to ~7000 (obs/sec).

Figure 17 presents the prediction time on the NWPU_RESISC45 dataset. This figure shows that the TNN classifier is the faster classifier of the listed classifiers as it has the lowest prediction speed, which is ~4200 (obs/sec) on the proposed 4Encoder-based ViT extracted features. In contrast, the WNN classifier has the longest time to calculate the prediction, which is ~5700 (obs/sec) on fused network weights. To observe the whole scenario, it is noted the proposed fused networks have a lower prediction speed as compared to the other networks due to their extracted excessive feature size, which is $N \times 1024$.

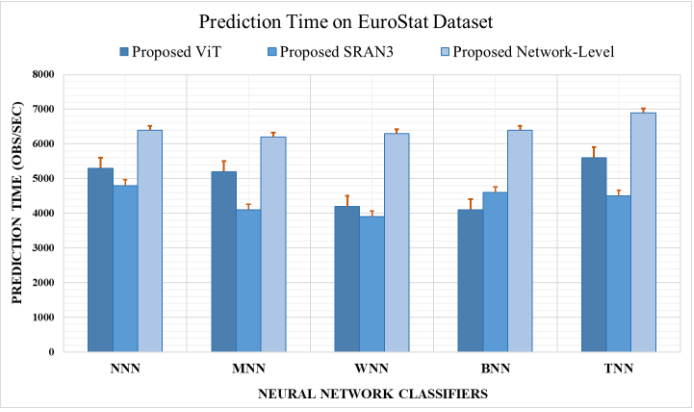


Figure 16: Prediction time analysis based on neural network classifiers for EuroSAT dataset

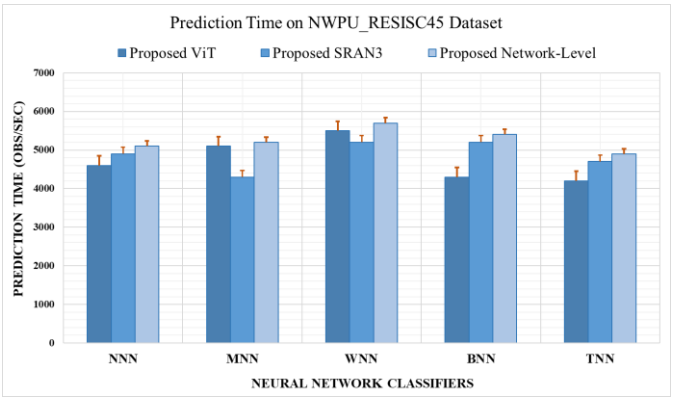


Figure 17: Prediction time analysis based on neural network classifiers using NWPU_RESISC45 dataset

2) Ablation Studies

In this section, detailed ablation studies have been conducted to validate the proposed architecture. The proposed stacked residual self-attention network (SRAN3) is evaluated with and without a residual connection by employing ADAM, SGDM, and RMSprop optimizers with various epochs, as shown in Table 4. This table shows that the SAN3 network achieved a % training accuracy of 90.4% by employing the SGDM optimizer and 100 epochs on the EuroSAT dataset. However, the SRAN3 model achieved a % training accuracy of 91.5% with the same configuration on the EuroSAT dataset. Moreover, both models are also evaluated on the NEPU_RESISC45 dataset, and it is perceived that the SAN3 model achieved 91.8% training accuracy when the network completes 100 epochs. The SRAN3 network gained a training accuracy of 93.4% with the SGDM optimizer. The purpose of implementing both architectures is to observe the impact of residual connection in the self-attention mechanism, and it has been proved that the residual connection significantly increased the training performance on the same configurations.

Table 4: Performance comparison among the self-attention and residual self-attention SRAN3 models based on various optimizers

Eurosat									
Optimizer	**Adam**			**SGDM**			**RMSprop**		
Models/Epochs	**30**	**70**	**100**	**30**	**70**	**100**	**30**	**70**	**100**
Self-Attention (SAN3)	82.1	82.6	85.9	83.9	88.7	**90.4**	83.0	85.4	88.1
Residual Self-Attention (SRAN3)	83.2	84.6	89.3	84.3	89.1	**91.5**	83.6	87.7	89.0
NWPU_RESISC45									
Self-Attention (SAN3)	85.6	86.5	89.7	87.4	89.2	**91.8**	86.1	86.9	88.6
Residual Self-Attention (SRAN3)	87.1	87.8	88.4	89.6	91.8	**93.4**	88.3	90.1	92.4

In the next phase, the proposed models are compared with the deep state-of-the-art networks based on basic configurations, including the number of layers, parameters, and models in megabytes. Table 5 shows that the DenseNet201 has 708 total layers, 20M parameters, and 77MB in size. The other pre-trained models like ResNet101, NasNetLarge, and GoogleNet have 347, 1243, and 144 layers

and 44.6M, 88.9M, 6.9M, parameters, and 167, 332, 27, 22.6 MB in size, respectively. Our proposed models 3Encoder ViT, SRAN3, and the fused network have 143, 86, and 139 layers, 5.7, 7.1, and 33 Million parameters, and 22.6 MB, 12.4 MB, and 115MB in size. The proposed models have fewer layers from the DensNet201, ResNet101, NasNetLarge, and GoogleNet. Moreover, the proposed fused network has fewer parameters from the NasNetLarge and ResNet101. The proposed SRAN3 is the smallest network in terms of size from the listed models and performed well for remote sensing datasets.

Table 5: Comparison of proposed fused and individual deep models with pre-trained CNN architectures based on the number of layers, parameters, and total size (in MB).

Models	No. of Layers	No. of parameters	Size in MB
DensNet201**[49]**	708	20 M	77 MB
ResNet101**[50]**	347	44.6 M	167 MB
NasNetLarge**[51]**	1243	88.9 M	332 MB
GoogleNet**[52]**	144	6.9 M	27 MB
Proposed 4Encoder ViT	143	5.7 M	22.6 MB
Proposed SRAN3	86	7.1 M	12.4 MB
Proposed Fused Network	139	33 M	115 MB

3) Comparisons

In the comparison section, we comprehensively compared several pre-trained models and state-of-the-art (SOTA) techniques. The comparison between the proposed architecture and pre-trained models is presented in Figure 18. This figure shows that the proposed fused network achieved the highest validation accuracy of 95.8% on the EuroSAT dataset, and the second highest validation accuracy is gained by the proposed 4encoder ViT, which is 93.7%. The proposed fused network again achieved a higher validation accuracy of 94.4% on the NWPU dataset. Also, the SRAN3 obtained the second-highest accuracy of 93.4% using the NWPU dataset.

Table 6 presents a comparison of the proposed architecture with SOTA techniques. Albarakati et al. [45] Xie et al. used the CNN technique to validate the NWPU dataset, obtaining an accuracy of 92%. [53] used the NWPU dataset and obtained the highest accuracy of 93.60%. Rubab et al. [54] introduced a fused CNN architecture and achieved an accuracy of 97.8% using the Eurostat dataset. Temenos et al. [55] Also, the EuroSAT dataset was used and obtained an accuracy of 94.72%. The proposed fused CNN architecture improved the accuracy of 98.4 and 95.7% for the EuroSAT and NWPU datasets, respectively. For the interpretation of the proposed model, we employed lime analysis and deep feature visualization, as shown in Figure 19. In the lime visualization, the brighter regions, such as yellow, red, and brown, strongly influence the model's prediction. In comparison, color regions, such as blue and green, have a weaker influence on the prediction.

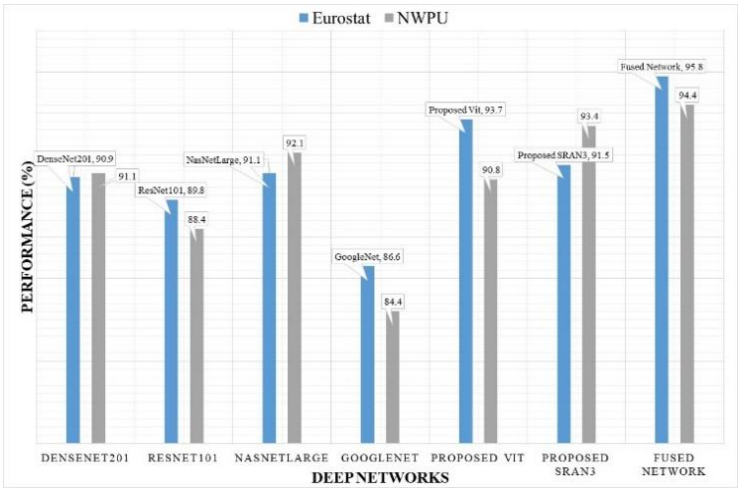


Figure 18: Proposed models comparison with state-of-the-art deep networks

Table 6: Comparison of proposed architecture accuracy with some SOTA techniques

Reference	Year	Dataset	Accuracy (%)
Albarakati et al. [45]	2024	NWPU	92.0
Xie et al. [53]	2021	NWPU	93.60
Rubab et al. [54]	2024	EuroSAT	97.8
Temenos et al. [55]	2023	EuroSAT	94.72
Proposed		EuroSAT	**98.4**
		NWPU	**95.7**

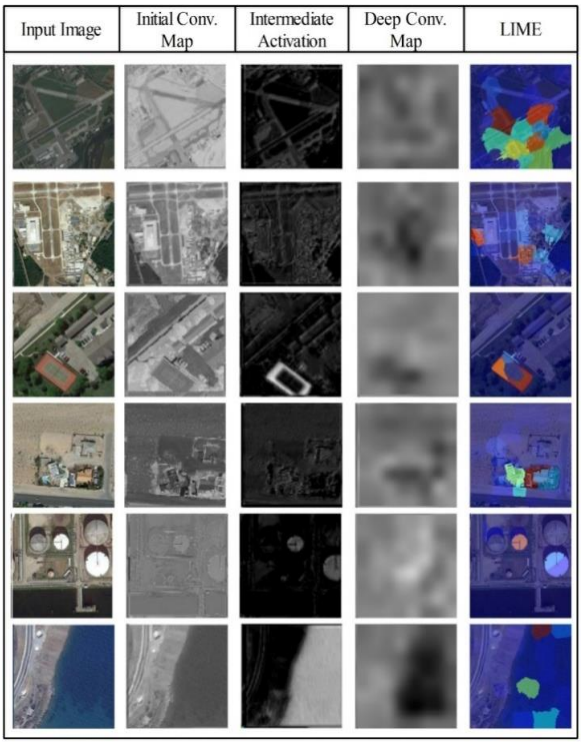


Figure 19: Feature visualization and Lime explanation

IV. Conclusion

This work proposes a novel network-level fused CNN architecture for LULC classification using remote sensing images. The dataset augmentation is performed based on statistical and high-resolution techniques. Two CNN architectures, SRAN3 and LViT-4E, are proposed, which are later fused using a depth-concatenation layer. Hyperparameters are selected using BO, and training is performed. In the testing phase, the trained model is tested and extracted features from the depth-concatenation layer. The extracted features are passed to the neural network classifiers, and the final prediction performance is obtained. The experimental process was conducted on two publically available datasets, EuroSAT and NWPU, and improved accuracy was obtained at 98.4 and 95.7%, respectively. Overall, we conclude with the following points:

- The dataset augmentation process improved the proposed architecture learning, which in turn improved the training, validation, and testing accuracy.
- The design of SRAN3 CNN architecture extracted the LULC area and performed more accurate predictions than the pre-trained CNN architectures.
- The proposed LViT-4E reduced the overall number of learnables later integrated with SRAN3 using a depth concatenation layer for improved accuracy and precision rate on the selected remote sensing dataset.
- The fused model is interpreted by the XAI technique and, in output, obtained improved prediction accuracy.
- The fusion of SRAN3 and LViT-4E CNN architecture improved accuracy and enhanced the overall prediction rate but increased the computational time, which is a dark side of this work.

The proposed model has outperformed the high-resolution samples. It decreases the performance of the raw samples, which is a limitation of the proposed model. In the future, we will design a framework-based CNN-MoE that handles the multi-resolution images, and we will improve the overall performance of the proposed model.

XI. References

[1] A. Khan, S. Gupta, and S. K. Gupta, "Multi-hazard disaster studies: Monitoring, detection, recovery, and management, based on emerging technologies and optimal techniques," *International journal of disaster risk reduction*, vol. 47, p. 101642, 2020.

[2] L. Zhu, T. Cui, A. Runa, X. Pan, W. Zhao, J. Xiang, *et al.*, "Robust remote sensing retrieval of key eutrophication indicators in coastal waters based on explainable machine learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 211, pp. 262-280, 2024.

[3] D. C. Marvin, L. P. Koh, A. J. Lynam, S. Wich, A. B. Davies, R. Krishnamurthy, *et al.*, "Integrating technologies for scalable ecology and conservation," *Global Ecology and Conservation*, vol. 7, pp. 262-275, 2016.

[4] Y. Xue, Y. Li, J. Guang, X. Zhang, and J. Guo, "Small satellite remote sensing and applications–history, current and future," *International journal of remote sensing*, vol. 29, pp. 4339-4372, 2008.

[5] J. Peng, J. Li, T. C. Ingalls, S. R. Schill, H. R. Kerner, and G. P. Asner, "A novel deep learning algorithm for broad scale seagrass extent mapping in shallow coastal environments," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 220, pp. 277-294, 2025.

[6] R. A. Schowengerdt, *Remote sensing: models and methods for image processing*: elsevier, 2006.

[7] J. B. Campbell and R. H. Wynne, *Introduction to remote sensing*: Guilford press, 2011.

[8] S. Khorram, F. H. Koch, C. F. Van der Wiele, and S. A. Nelson, *Remote sensing*: Springer Science & Business Media, 2012.

[9] S. Baek and W. Kim, "Review on Hyperspectral Remote Sensing of Tidal Zones," *Ocean Science Journal*, vol. 60, pp. 1-21, 2025.

[10] P. H. Swain, "Pattern recognition: a basis for remote sensing data analysis," 1973.

[11] D. G. Leckie, "Advances in remote sensing technologies for forest surveys and management," *Canadian Journal of Forest Research*, vol. 20, pp. 464-483, 1990.

[12] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22-36, 2016.

[13] A. Shah, M. Attique Khan, A. Ibrahim Alzahrani, N. Alalwan, A. Hamza, S. Manic, *et al.*, "FuzzyShallow: A framework of deep shallow neural networks and modified tree growth optimization for agriculture land cover and fruit disease recognition from remote sensing and digital imaging," *Measurement*, vol. 237, p. 115224, 2024/09/30/ 2024.

[14] A. Hamza, M. A. Khan, S. u. Rehman, M. Al-Khalidi, A. I. Alzahrani, N. Alalwan, *et al.*, "A Novel Bottleneck Residual and Self-Attention Fusion-Assisted Architecture for Land Use Recognition in Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 2995-3009, 2024.

[15] R. P. Gupta, *Remote sensing geology*: Springer, 2017.

[16] W. Musakwa and A. Van Niekerk, "Earth observation for sustainable urban planning in developing countries: needs, trends, and future directions," *Journal of Planning Literature*, vol. 30, pp. 149-160, 2015.

[17] P. Scala, G. Manno, and G. Ciraolo, "Semantic segmentation of coastal aerial/satellite images using deep learning techniques: An application to coastline detection," *Computers & Geosciences*, vol. 192, p. 105704, 2024.

[18] I. Haider, M. A. Khan, M. Nazir, A. Hamza, O. Alqahtani, M. T. H. Alouane, *et al.*, "Crops Leaf Disease Recognition From Digital and RS Imaging Using Fusion of Multi Self-Attention RBNet Deep Architectures and Modified Dragonfly Optimization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 7260-7277, 2024.

[19] J. Yuan, L. Wang, T. Wang, A. K. Bashir, M. M. Al Dabel, J. Wang, *et al.*, "YOLOv8-RD: High-Robust Pine Wilt Disease Detection Method Based on Residual Fuzzy YOLOv8," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[20] M. A. Moharram and D. M. Sundaram, "Land use and land cover classification with hyperspectral data: A comprehensive review of methods, challenges and future directions," *Neurocomputing*, vol. 536, pp. 90-113, 2023.

[21] L. Wang, J. Cai, T. Wang, J. Zhao, T. R. Gadekallu, and K. Fang, "Pine wilt disease detection based on uav remote sensing with an improved yolo model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[22] L. Lopez-Fuentes, J. van de Weijer, M. González-Hidalgo, H. Skinnemoen, and A. D. Bagdanov, "Review on computer vision techniques in emergency situations," *Multimedia Tools and Applications*, vol. 77, pp. 17069-17107, 2018.

[23] C. Atzberger, "Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs," *Remote sensing*, vol. 5, pp. 949-981, 2013.

[24] M. Digra, R. Dhir, and N. Sharma, "Land use land cover classification of remote sensing images based on the deep learning approaches: a statistical analysis and review," *Arabian Journal of Geosciences*, vol. 15, p. 1003, 2022.

[25] M. Mazzei, "An unsupervised machine learning approach in remote sensing data," in *Computational Science and Its Applications–ICCSA 2019: 19th International Conference, Saint Petersburg, Russia, July 1–4, 2019, Proceedings, Part III 19*, 2019, pp. 435-447.

[26] H. Feng, Q. Li, W. Wang, A. K. Bashir, A. K. Singh, J. Xu, *et al.*, "Security of target recognition for UAV forestry remote sensing based on multi-source data fusion transformer framework," *Information Fusion*, vol. 112, p. 102555, 2024.

[27] J. Fan, M. Han, and J. Wang, "Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation," *Pattern Recognition*, vol. 42, pp. 2527-2540, 2009.

[28] J. Dong, W. Perrizo, Q. Ding, and J. Zhou, "The application of association rule mining to remotely sensed data," in *Proceedings of the 2000 ACM symposium on Applied computing-Volume 1*, 2000, pp. 340-345.

[29] A. W. Abbas, N. Minallh, N. Ahmad, S. A. R. Abid, and M. A. A. Khan, "K-Means and ISODATA clustering algorithms for landcover classification using remote sensing," *Sindh University Research Journal-SURJ (Science Series)*, vol. 48, 2016.

[30] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, pp. 1735-1739, 2017.

[31] M. A. Savelonas, C. N. Veinidis, and T. K. Bartsokas, "Computer Vision and Pattern Recognition for the Analysis of 2D/3D Remote Sensing Data in Geoscience: A Survey," *Remote Sensing*, vol. 14, p. 6017, 2022.

[32] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu, "Feature fusion: parallel strategy vs. serial strategy," *Pattern recognition*, vol. 36, pp. 1369-1381, 2003.

[33] B. Victor, Z. He, and A. Nibali, "A systematic review of the use of Deep Learning in Satellite Imagery for Agriculture," *arXiv preprint arXiv:2210.01272*, 2022.

[34] X. Lei, H. Pan, and X. Huang, "A dilated CNN model for image classification," *IEEE Access*, vol. 7, pp. 124087-124095, 2019.

[35] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sensing*, vol. 10, p. 144, 2018.

[36] P. Deepan and L. Sudha, "Fusion of deep learning models for improving classification accuracy of remote sensing images," *Journal of mechanics of continua and mathematical sciences*, vol. 14, pp. 189-201, 2019.

[37] M. Pritt and G. Chern, "Satellite image classification with deep learning," in *2017 IEEE applied imagery pattern recognition workshop (AIPR)*, 2017, pp. 1-7.

[38] K. Gao, M. Chen, S. Narges Fatholahi, H. He, H. Xu, J. Marcato Junior, *et al.*, "A region-based deep learning approach to instance segmentation of aerial orthoimagery for building rooftop extraction," *Geomatica*, vol. 75, pp. 148-164, 2022.

[39] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Asian conference on computer vision*, 2016, pp. 180-196.

[40] K. Vani, "Deep learning based forest fire classification and detection in satellite images," in *2019 11th international conference on advanced computing (ICoAC)*, 2019, pp. 61-65.

[41] C. Zhang, Y. Cui, Z. Zhu, S. Jiang, and W. Jiang, "Building height extraction from GF-7 satellite images based on roof contour constrained stereo matching," *Remote sensing*, vol. 14, p. 1566, 2022.

[42] M. K. Hasan, S. Islam, T. R. Gadekallu, A. F. Ismail, S. Amanlou, and S. N. H. S. Abdullah, "Novel EBBDSA based Resource Allocation Technique for Interference Mitigation in 5G Heterogeneous Network," *Computer Communications*, vol. 209, pp. 320-330, 2023.

[43] J. Chen, C. Wang, Z. Ma, J. Chen, D. He, and S. Ackland, "Remote sensing scene classification based on convolutional neural networks pre-trained using attention-guided sparse filters," *Remote Sensing*, vol. 10, p. 290, 2018.

[44] S. S. Alam, S. Chakraborty, F. C. Jain, S. Deb, R. Singh, and D. C. Weindorf, "Proximal sensor integration for land use classification and soil analysis in a coastal environment," *Case Studies in Chemical and Environmental Engineering*, vol. 11, p. 101079, 2025.

[45] H. M. Albarakati, S. ur Rehman, M. A. Khan, A. Hamza, J. Aftab, A. Alasiry, *et al.*, "A Unified Super-Resolution Framework of Remote Sensing Satellite Images Classification based on Information Fusion of Novel Deep Convolutional Neural Network Architectures," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[46] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, pp. 2217-2226, 2019.

[47] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection," *Biomedical Signal Processing and Control*, vol. 82, p. 104561, 2023.

[48] M. Santos, "Bayesian Optimization for Hyperparameter Tuning," *Journal of Bioinformatics and Artificial Intelligence*, vol. 2, pp. 1-13, 2022.

[49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[51] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697-8710.

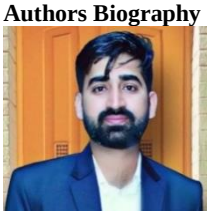
[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[53] H. Xie, Y. Chen, and P. Ghamisi, "Remote sensing image scene classification via label augmentation and intra-class constraint," *Remote Sensing*, vol. 13, p. 2566, 2021.

[54] S. Rubab, M. A. Khan, A. Hamza, H. M. Albarakati, O. Saidani, A. Alshardan, *et al.*, "A novel network level fusion architecture of proposed self-attention and vision transformer models for land use and land cover classification from remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[55] A. Temenos, N. Temenos, M. Kaselimi, A. Doulamis, and N. Doulamis, "Interpretable deep learning framework for land use and land cover classification in remote sensing using SHAP," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1-5, 2023.

Authors Biography



Muhammad Kashif Bhatti received the bachelor's degree in 2016 from Arid Agriculture, Rawalpindi, Pakistan, and the master's degree from the University of Lahore, Lahore, Pakistan, in 2018, both in computer science. He is currently working toward the Ph.D. degree in the area of remote sensing with HITEC University, Taxila, Pakistan. He is the coauthor of few reputed journal papers in the area of remote sensing.

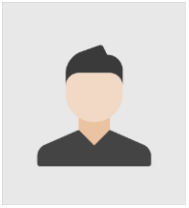
Saima Shaheen received the Doctoral degree in Software Engineering from National University of Sciences and Technology, Islamabad. Currently, She is assistant professor of HITEC University, Taxila, Pakistan. Her research interest including computer networks, Multimedia Computation, and computerized software models.

Muhammad Attique Khan (Member IEEE) received the master's and Ph.D. degrees in human activity recognition for application of video surveillance and skin lesion classification using deep learning from COMSATS University Islamabad, Islamabad, Pakistan, in 2018 and 2022, respectively. He is currently an Assistant Professor with AI Department, Prince Mohammad Bin Fahd, Al-Khobar, Saudi Arabia. His primary research focus in recent years is medical imaging, COVID-19, MRI analysis, video surveillance, human gait recognition, and agriculture plants using deep learning. He has above 350 publications that have more than 16 000+ citations and an impact factor of 1050+ with h-index 74 and i-index 230. He is the Reviewer of several reputed journals, such as the IEEE Transaction on Industrial Informatics, IEEE Transaction of Neural Networks, Pattern Recognition Letters, Multimedia Tools and Application, Computers and Electronics in Agriculture, IET Image Processing, Biomedical Signal Processing Control, IET Computer Vision, EURASIP Journal of Image and Video Processing, IEEE Access, MDPI Sensors, MDPI Electronics, MDPI Applied Sciences, MDPI Diagnostics, and MDPI Cancers.

Ameer Hamza is currently working toward the Ph.D. degree in computer science with KTU University, Kaunas, Lithuania. His major interests include object detection and recognition, video surveillance, medical, and agriculture using deep learning and machine learning. He has published 20 impact factor papers to date.

Dr. Ali Arishi is an assistant professor in the Industrial Engineering department at King Khalid University. He earned his Ph.D. in Industrial Engineering from Wichita State University, and he holds an M.S. from Clemson University, having completed his B.S. at King Khalid University. Ali's research interests center on the applications of artificial intelligence in supply chain management, particularly in engineering contexts. He has published in well-respected journals that focus on the engineering applications of AI, contributing valuable insights to the advancement of the field.

Dina Abdulaziz AlHammadi received the Ph.D. degree in computer science from the University of Sheffield, Sheffield, U.K., in 2021.,She is currently an Assistant Professor with the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, specializing in artificial intelligence. She is actively involved in mentoring students and collaborating with industry partners to bridge the gap between academia and practical applications of AI. Her research interests include AI, human-computer interaction, personality recognition, and social AI.



Shabbab Ali Algamdi has served as a faculty member in the software engineering department at Prince Sattam bin Abdulaziz University, where I teach courses related to the software engineering department. My research interests include artificial intelligence, deep learning (particularly in natural language processing and data science), and the usability discipline of human-computer interaction. I earned my master’s degree in software engineering from Southern Methodist University (SMU) and my PhD in computer science from the University of North Texas. Additionally, I am a member of the ACM and IEEE organizations.

Yunyoung Nam (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, South Korea, in 2001, 2003, and 2007, respectively. He was a Senior Researcher with the Center of Excellence in Ubiquitous System, Stony Brook University, Stony Brook, NY, USA, from 2007 to 2010, where he was a Postdoctoral Researcher, from 2009 to 2013. He was a Research Professor with Ajou University, from 2010 to 2011. He was a Postdoctoral Fellow with Worcester Polytechnic Institute, Worcester, MA, USA, from 2013 to 2014. He was the Director of the ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, from 2017 to 2020. He has been the Director of the ICT Convergence Research Center, Soonchunhyang University, since 2020, where he is currently an Assistant Professor with the Department of Computer Science and Engineering. His research interests include multimedia database, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare systems.