A Novel Deep Learning Architecture for Agriculture Land Cover and Land Use Classification from Remote Sensing Images Based on Network-Level Fusion of Self-Attention Architecture

Hussain Mobarak Albarakati ⬡, Muhammad Attique Khan ⬡, *Member, IEEE*, Ameer Hamza ⬡, Faheem Khan ⬡, Naoufel Kraiem ⬡, Leila Jamel ⬡, Latifah Almuqren ⬡, and Roobaea Alroobaea ⬡

Abstract—AI-driven precision agriculture applications can benefit from the large data source that remote sensing (RS) provides, as it can gather agricultural monitoring data at various scales throughout the year. Numerous advantages for sustainable agricultural applications, including yield prediction, crop monitoring, and climate change adaptation, can be obtained from RS and artificial intelligence. In this work, we proposed a fully automated optimized self-attention fused convolutional neural network (CNN) architecture for land use and land cover classification using RS data. A new contrast enhancement equation has been proposed and utilized in the proposed architecture for the data augmentation. After that, a fused self-attention CNN architecture was proposed. The proposed architecture initially consists of two custom models named IBNR-65 and Densenet-64. Both models have been designed based on the inverted bottleneck residual mechanism and dense blocks. After that, both models were fused using a depth-wise concatenation and append a self-attention layer for deep features extraction. After that, we trained the model and performed classification using neural network (NN) classifiers. The results obtained from the NN classifiers are insufficient; therefore, we implemented a Bayesian

optimization and fine-tuned the hyperparameters of NN. In addition, we proposed a quantum hippopotamus optimization algorithm for the best feature selection. The selected features are finally classified using fine-tuned NN classifiers and obtained improved accuracy of 98.20, 89.50, and 91.70%, and the highest precision rate is 98.23, recall is 98.20, and F1-score is 98.21, respectively, for SIRI-WHU, EuroSAT, and NWPU datasets. Moreover, a detailed ablation study was conducted, and the performance was compared with SOTA. The proposed architecture shows improved accuracy, sensitivity, precision, and computational time performance.

Index Terms—Agriculture, data augmentation, deep learning, optimization, remote sensing (RS), self-attention.

I. INTRODUCTION

ADVANCES in remote sensing (RS) technology have made it possible to obtain a significant amount of satellite data quickly. For RS community researchers, new challenges are always opened based on the high-resolution satellite images [1]. Computer vision researchers have extensively utilized RS images for many semantic tasks, including but not limited to road segmentation, building extraction, land cover classification, IOT, and agricultural land classification [2], [3], [4]. The land cover classification achieved remarkable attention in computer vision due to its essential applications such as urban planning, crop fields, and landslide hazards [3]. The RS data are not easy to use because several things fall under the same category and might be seen in the same image, i.e., the vegetation category includes forest regions, herbaceous plants, and permanent crops [5]. A few sample images are illustrated in Fig. 1.

Several techniques were introduced in the literature to classify land cover from RS images [6], [7]. The presented techniques are based on supervised learning and unsupervised learning. In unsupervised learning, clustering techniques have usually been employed for classification [8], [9], such as K-means and fuzzy C-means; however, these techniques are not suitable and efficient due to many labeled imaging RS images [10]. The supervised learning methods used traditional methods such as handcrafted features and classification using machine learning classifiers. The handcrafted features are extracted based on prior information such as texture, shape, and point features [11]. However, for complex scenes, it is not easy to extract the most discriminative

Hussain Mobarak Albarakati is with the Computer Engineering and Network Department, College of Computer and Information Systems, Umm Al-Qura University, Makkah 24382, Saudi Arabia (e-mail: hmbarkti@uqu.edu.sa).

Muhammad Attique Khan is with the Department of Computer Science and Mathematics, Lebanese American University, Beirut 13-5053, Lebanon, and also with the Department of CS, HITEC University, Taxila 47080, Pakistan.

Ameer Hamza is with the Department of CS, HITEC University, Taxila 47080, Pakistan.

Faheem Khan is with the Gachon University, Seongnam 13120, South Korea (e-mail: faheem@gachon.ac.kr).

Naoufel Kraiem is with the College of Computer Science, King Khalid University, Abha, SA 61421, Saudi Arabia (e-mail: nkraiem@kku.edu.sa).

Leila Jamel and Latifah Almuqren are with the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia (e-mail: Lmjamel@pnu.edu.sa; laAlmuqren@pnu.edu.sa).

Roobaea Alroobaea is with the Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia.

Digital Object Identifier 10.1109/JSTARS.2024.3369950

Fig. 1. Few sample images of land cover RS.

features [12], [13]. Feature selection is an important area in pattern recognition research, and many techniques have been introduced in the literature. Feature selection techniques mainly aim to reduce irrelevant information from the original feature space and minimize the computation time [14].

Deep learning has gained huge success worldwide for several applications, especially RS and object classification [15]. Deep learning is famous due to its large learning capacity and better performance of huge datasets. Convolutional neural networks (CNNs), a widely used deep learning model, can learn abstract features from images layer-by-layer, and it is utilized in diverse fields like healthcare [16], action recognition [17], satellite imaging [18], fraud detection [19], and many more. A simple CNN architecture consists of several intermediate layers, such as a convolutional layer, a pooling layer, a ReLu activation layer, a batch-normalization layer, an additional layer, a fully connected (FC) layer, and a softmax layer [20]. Several recent studies used pretrained deep learning models for land cover classification using RS images. Papoutsis et al. [21] introduced a CNN architecture for land cover classification from RS images. The presented architecture is based on multilayer perceptron and vision transformers.

Moreover, they added an EfficientNet mechanism to improve the training time and accuracy. They also compared the presented architecture with baseline ResNet50, showing improved accuracy and time. Ma et al. [22] presented a features enhancement neural network (NN) for land cover classification. They also added a self-attention module to extract the local information of the images and reduce the classification loss. Compared with the previous model named PSPNet, this network improved accuracy by 2%. For the crop classification task from the RS images, Patel et al. [23] presented a comparative study that focused on the pretrained model's performance, such as VGG16, VGG19, ResNet, and DenseNet. They trained these models on the RS image data and analyzed the performance. In addition, they analyze the performance of 2D CNN and 3D CNN custom architectures. Based on the results, they conclude that the custom models show better performance. Helber et al. [24] introduced a novel RS dataset named EuroSAT for the land cover classification task. The dataset that was produced was tested on a deep learning architecture, and improved classification accuracy was obtained.

Kussul et al. [25] introduced a deep learning technique for land use and crop type classification using RS images. A review study is also presented by Vali et al. [26] for the land cover classification techniques. This study discussed the preprocessing, features engineering, and classification techniques for the land cover classification. Zhang et al. [27] presented a joint Deep CNN architecture using RS data for the land cover classification. They combined the patch-based CNN and pixel-based MLP with joint reinforcement to improve classification accuracy. Otal et al. [28] presented a framework for postbuyout land cover mapping using harnessing deep learning. They used FEMA's and grant program (HMGP) buyout dataset for experimental purpose. They gathered 40 053 satellite images of buyout lands and they employed deep learning models to evaluate the performance and they achieved 98.86% AUC score. Temenos et al. [5] presented a deep learning-based interpretable framework for classifying land use and land cover using satellite images. The authors use the EuroSAT dataset and employ customized CNNs. They achieved 94.72% accuracy and they applied SHAP to verify the performance of models. Vinaykumar et al. [29] presented a hybrid deep learning and optimal guidance whale optimization for land cover classification using satellite images. In this article, the authors applied alexnet and resnet50 for feature extraction and employed whale optimization for best feature selection. The authors used bi-lstm for the classification purpose. They achieved 89.57% on AID and 93.21% on NWHP dataset. A few more studies are also introduced, such as the high-resolution adaptive model [30], 3D-HRNET [31], crop classification [32], and the Northern Border Region technique [33].

Although the techniques mentioned earlier show impressive performance, there is room for improvement in accuracy and computational time optimization. The fusion process has shown an impact in the last few years for the improvement of accuracy; however, there is a drawback of this step, which is an increase in computational time. In the fusion process, the researchers usually fuse the features of two models [34]; however, this process is inefficient and consumes more time. To overcome this problem, a few researchers employed feature selection techniques that remove the irrelevant features from the final classification. In addition to that, the performance has been improved significantly after employing the optimization process [35], [36]. A few latest studies also focused on deep learning for the agri-yield classification using RS images [20], [37], [38], [39], [40], [41]. Overall, it is a complex process that first fuses the features of both models and then selects the best of them; therefore, it is important to consider a more optimized approach, such as network-level fusion.

In the network-level fusion, two models are created and combined within the network for improved classification accuracy and less computational cost. However, controlling the total learnable when performing a fusion process within a network is challenging. To tackle this challenge, we proposed a novel CNN architecture based on network-level fusion and optimization in this work. In the fusion process, we design two custom networks based on the bottleneck and dense mechanisms and then concatenate using a depth concatenation layer. After the fusion layer, a self-attention layer has been added to extract the local information of the image. Further, the model performance
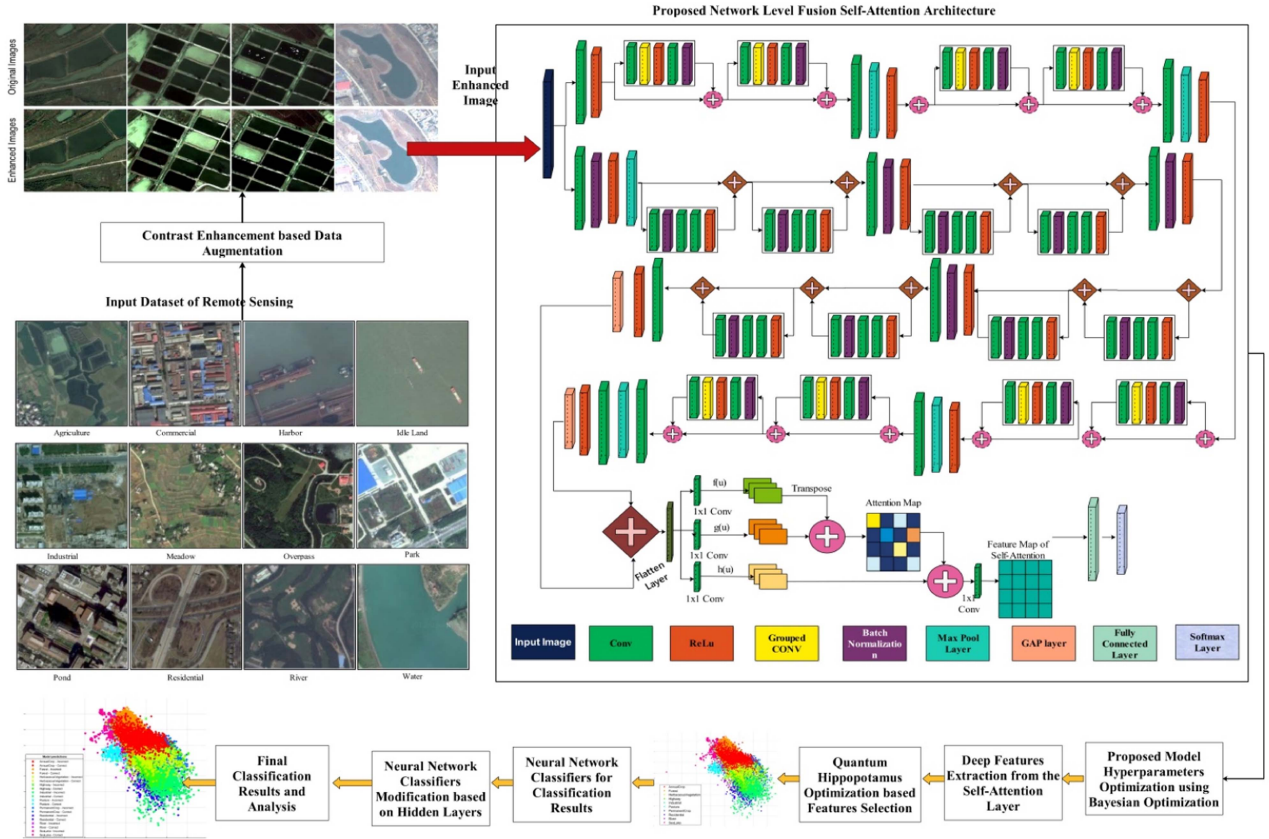
Fig. 2. Proposed framework of land cover classification from RS images.

has been optimized using a Bayesian optimization (BO) and quantum hippopotamus optimization algorithm (QHPO). Our major contributions of this work are given as follows.

- To address the problem of an imbalanced dataset, we performed data augmentation at the initial step using the contrast stretching technique and employed the updated dataset for the training process.
- The samples of satellite images have complex patterns and are difficult to recognize; therefore, we proposed a novel network-level fusion Self-Attention CNN architecture. The proposed architecture's upper part is based on two mechanisms bottleneck and dense blocks. The depth concatenation layer combined both architectures and added a Self-Attention layer for local information extraction that followed the FC and softmax layers.
- A BO technique has been implemented using an expected improvement (EI) acquisition function that optimizes the hyperparameters of the proposed model instead of past knowledge selection.
- The irrelevant and redundant features misclassified the RS classes; therefore, we proposed a new QHPO that selects the best features and reduces the computation time.

The rest of this article is organized into four main sections. Section II presented a detailed study of the proposed methodology, including dataset description, data augmentation, proposed CNN-based self-attention model, BO learning, and QHPO-bases feature selection. In Section III detailed results of the proposed framework are discussed. Section IV concludes the overall proposed research.

II. PROPOSED METHODOLOGY

The proposed land cover classification framework has been presented in this section. The proposed framework comprises a novel CNN self-attention fused architecture for feature extraction. Fig. 2 illustrates the proposed framework of land cover classification from RS images. In the proposed framework, data augmentation has been performed at the initial step using a hybrid contrast enhancement technique. After that, fused self-attention deep learning architecture was proposed, and the hyperparameters used BO for the training process. Deep features have been extracted from the self-attention layer and optimized using a novel QHPO. The proposed QHPO selects the best features for final classification using NN classifiers.

A. Dataset Description

In this work, we employed three datasets for the classification purposes such as EuroSAT [24], NWPU-RESISC45 [42], and SIRI-WHU [43]. These datasets are publicly available for the research purposes of the RS domain. The nature of the selected dataset is RGB. A brief description of each dataset is given in the following.

The *EuroSAT dataset* consists of 10 classes, as summarized in Table I. The aim of this dataset was land cover and land use

TABLE I
DESCRIPTION OF SELECTED DATASETS FOR THE CLASSIFICATION OF LAND
COVER AND LAND USE FROM RS IMAGES

EuroSAT		NWPU-RESISC45		SIRI-WHU	
Class name	# Images	Class name	# Images	Class name	# Images
AnnualCrop	3000	Airfield	1400	Agriculture	200
Forest	3000	Anchorage	700	Commercial	200
HerbaceousVegetation	3000	Beach	700	Harbor	200
Highway	3000	Dense residential	700	idle land	200
Industrial	2500	Farm	1400	Industrial	200
Pasture	2000	Flyover	700	Meadow	200
PermanentCrop	2500	Forest	700	Overpass	200
Residential	3000	Game space	1400	Park	200
River	2500	Parking space	700	Pond	200
Sea Lake	3000	River	700	Residential	200
		Sparse residential	700	River	200
Total	**27500**	Storage cisterns	700	Water	200
		Total	**10500**	**Total**	**2400**

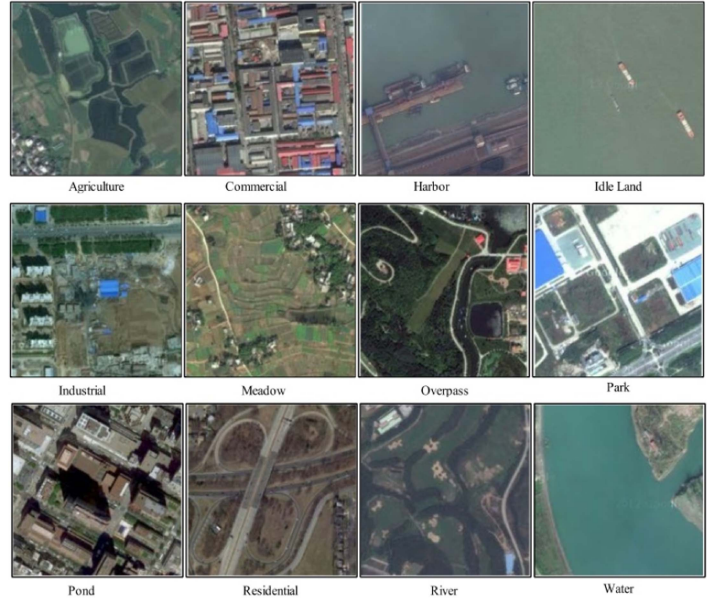Bold values shows the best results.

Fig. 3. Few sample RS images of SIRI-WHU dataset for land cover classification.
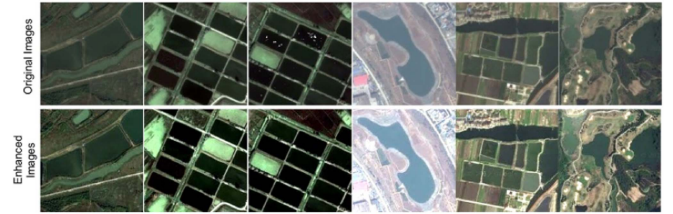
Fig. 4. Visual illustration of the proposed contrast enhancement approach for RS images.

classification. Each image of this dataset has 64 by 64 pixels and a 10-m ground sampling distance. Every one of them was gathered by the Sentinel-2 satellite. The total number of images in this dataset is 27 500 [24].

Northwestern Polytechnic University created the *NWPU dataset* in 2017 [42]. This dataset is publically available for the researcher of RS. There are 12 classes in this dataset; each class contains images in the range of 700 to 1400 with a pixel ratio of 256×256. Within the scene classes, the spatial resolution drops from 30.0 to 0.2 m per pixel. The dataset is quite difficult because of the rich image variations, some discrepancies within the scene classes, and some commonalities between the scene classes. Table I shows the number of images in each class.

The *SIRI-WHU dataset* is a publicly available RS database used for land use classification [43]. It consists of 12 classes and 2400 images (200 images in each class). Each image is 200×200 and has a spatial resolution of 2 m. Table I demonstrates the description of each class. Moreover, a few sample RS images are shown in Fig. 3.

B. Data Augmentation of Training Set

In this work, we utilized the contrast enhancement technique to generate new images instead of traditional rotate and flip operations, a few sample images are shown in Fig. 4. A new mathematical fitness function is proposed for contrast enhancement. Consider I is an input image of dimension $256 \times 256 \times 3$,

and the resultant image is denoted by F. A contrast function is implemented at the initial step to improve the image pixel intensity values as follows:

$$\tilde{\mathbb{F}}(u) = \frac{1}{\int_0^1 \mathbb{I}^{a-1}(1-\mathbb{I})^{b-1}d\mathbb{I}} \times \int_0^u \mathbb{I}^{a-1}(1-\mathbb{I})^{b-1}d\mathbb{I} \quad (1)$$

where $\mathbb{I}$ is an integrated image, a and b are two adjusted parameters that control the brightness of an image during the processing. Finally, the resultant $\tilde{\mathbb{F}}$ is passed to the fitness function and obtains the final enhanced image as follows:

$$\mathbb{G}\left(\tilde{\mathbb{F}}\right) = u_1 \log\left(\left(\log(S\left(\tilde{\mathbb{F}}\right))\right)\right) + u_2 H\left(\tilde{\mathbb{F}}\right)$$
$$+ u_3 \log\left(\mathbb{I} + \tilde{\mathbb{F}}\right) \quad (2)$$

where u_1, u_2, and u_3 are weight coefficients, $S(\tilde{\mathbb{F}})$ denotes the sum of the fringe intensities of the image, and $H(\tilde{\mathbb{F}})$ is entropy value. This technique is employed for only less number of images classes such as in EuroSAT, the augmentation is performed for industrial, pasture, permanent crop, and river class. In these classes, we generated more images until the total images are reached to 3000 (maximum images of this dataset). Similarly, this process is performed for the rest of the datasets.
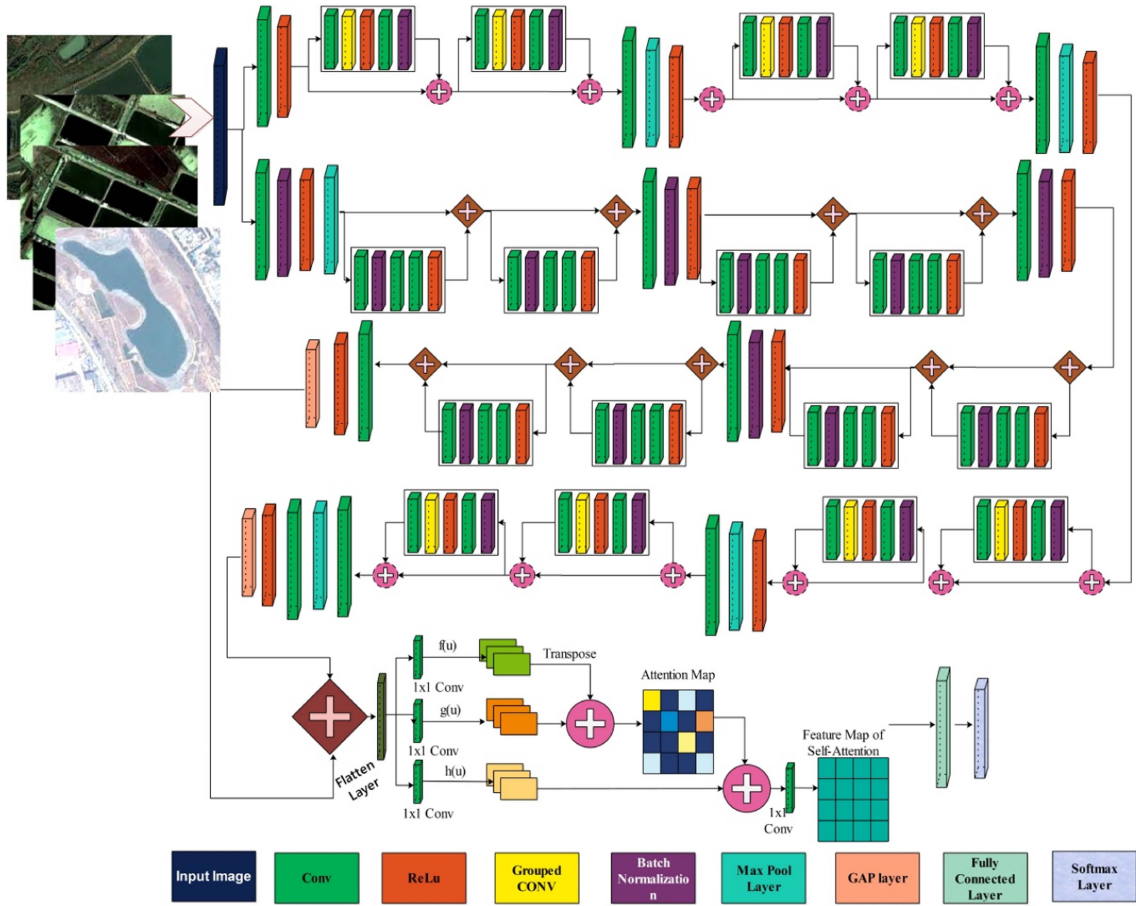
Fig. 5.   Proposed fused self-attention deep learning architecture for land cover and land use classification.

C. Proposed Methodology

In this work, we proposed a fused self-attention deep learning architecture for land cover and land use classification from RS images. The proposed architecture consists of two CNN architectures (bottleneck and density mechanism) fused in the last stage and then embedded in a self-attention layer for feature extraction. Each architecture consists of several intermediate layers: a convolutional layer, pooling layer, ReLu activation layer, batch normalization layer, self-attention layer, FC layer, and a softmax layer. A visual architecture of the proposed fused self-attention CNN architecture is illustrated in Fig. 5. This figure illustrates that the enhanced images have been employed as input to this network called the input layer. After that, two networks were designed: Inverted bottleneck residual 65 layered (IBNR-65) and Densenet-64. The IBNR-65 comprises 65 hidden layers in the inverted bottleneck fashion, whereas the Densenet-64 includes 64. In the IBNR-65 network, the initial convolutional layer has been added to convolve the features on depth size 32 and kernel size 3×3. Mathematically, the activation of the convolutional layer has been defined as follows:

$$\tilde{\phi}^{(i)}\left(l \right) = \max \left(0, \tilde{\beta}^{j(l)} + \sum_{j} J^{i,j(l)} * \mathbb{F}^{i(l)} \right) \quad (3)$$

where $\mathbb{F}^{i(l)}$ denotes the input activation map, $\tilde{\phi}^{(i)}(l)$ is ith output activation map, $\tilde{\beta}^{j(l)}$ is a bias of jth output activation map, and $J^{i,j(l)}$ is a convolutional kernel between ith and jth maps.

After the convolutional layer, a ReLu activation layer has been added to convert nonlinear separable data into linear form and then fed to the next layer as an input. Another important layer in this architecture is the max pooling layer. The purpose of this layer is to scale down the image but keep the important features for the recognition task. Mathematically, this layer is defined as follows:

$$\tilde{\phi}_{jk}^{(i)} = \max \left(\mathbb{F}_{js+m,ks+n}^{i} \right). \quad (4)$$

After this layer, a batch normalization layer is added to fasten the training process. In addition, the skip connections have been combined using an additional layer. The skip connections reduce the complexity of the proposed model and improve the efficiency based on the combined information as follows:

$$\tilde{\phi}_{skip}^{(i)} = L_i + CL \quad (5)$$

where L_i denotes the output layer of the respective path and CL denotes the skip layer that is connected with L_i using an additional layer.

We added several residual and dense blocks using these layers, as shown in Fig. 5, for both IBNR-65 and Densenet-64. Each block consists of a convolutional 2D layer, batch normalization
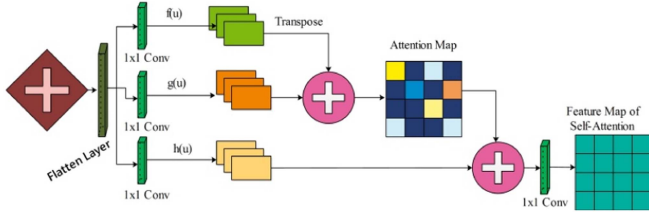
Fig. 6. Self-Attention module for local features extraction of the image.

layer, grouped convolutional layer, max-pooling layers, and ReLu activation. The depth size starts from 32 and ends with 1024. Finally, a global average pool layer has been added for both networks and depth-wise concatenated in a single layer named the depth concatenation layer. After the depth-wise concatenation, we added a self-attention layer to extract the most important information of the input image.

1) Self-Attention: Self-attention networks (SANs) are becoming increasingly popular because of their high computational parallelization and adaptability when modeling interdependence. In the CNN architecture, the self-attention module improves the performance of a network due to the attention on the more overriding area of the image. Through SAN, local features of the image have been extracted. Consider we have a depth concatenation layer features denoted by $\tilde{\phi} \in \mathbb{R}^{C \times N}$, where C denotes the number of channels. A 1×1 convolutional operation has been performed on $\tilde{\phi}$ and obtained three 1D matrix f, g, and h, as shown in Fig. 6. Mathematically, f, g, and h are defined as follows:

$$f\left(\tilde{\phi}\right) = \psi_f \tilde{\phi}, g\left(\tilde{\phi}\right) = \psi_g \tilde{\phi} \tag{6}$$

$$h\left(\tilde{\phi}\right) = \psi_h \tilde{\phi} \tag{7}$$

$$\psi_f, \psi_g, \psi_h \in \mathbb{R}^{C^* \times C}. \tag{8}$$

The feature maps of $f(\tilde{\phi})$ and $g(\tilde{\phi})$ are combined by performing a series of softmax.

$$\mathbb{S}_{ji} = \frac{\exp\left(\psi_{ij}\right)}{\sum_{i=1}^{N} \exp\left(\psi_{ij}\right)} \tag{9}$$

$$\psi_{ij} = f\left(\tilde{\phi}_i\right)^T g\left(\tilde{\phi}_j\right). \tag{10}$$

The final output of the self-attention layer is defined in (11), which is further utilized for the features extraction map.

$$O_j = \Phi\left(\sum_{i=1}^{N} \mathbb{S}_{ji} h\left(\tilde{\phi}_i\right)\right). \tag{11}$$

The resultant attention map feature matrix is fed to an FC layer. In the FC layer, all features are FC. Mathematically, it is defined as follows:

$$\tilde{\phi}_{jk}^i = A\left(w\right)\left(\sum_{i=1}^{n(l-1)} \tilde{\phi}^{l-1}\left(i\right) . \mathbb{X}^l\left(i,j\right) + \beta_i^{(l)}\right) \tag{12}$$

where $l-1$ denotes the number of neurons in the previous layer $(l-1)$, $\mathbb{X}^l(i,j)$ denotes the weight for the connection from a neuron i in layer $(l-1)$ to neuron j in layer l. The $A\left(w\right)$ denotes the activation function of this layer. The output of this layer is fed to the softmax layer for multiclass land cover and land use classification.

D. Training and Feature Extraction

The training was performed after the design of the proposed self-attention fused CNN architecture. Several hyperparameters have been employed in the training process, such as learning rate, mini-batch size, epochs, and optimizer. The values of these hyperparameters have been computed using BO. The model is trained on 50% of images that was splited before the augmentation process. Augmentation was performed on the split images. After a model's training, deep features are extracted from the self-attention layer with the size of $N \times 2048$ that is further used for classification. Many NN classifiers, such as narrow, medium, wide, trilayered, and bilayered, have been employed for the classification results. The accuracy and a few other measures have been computed for each classifier. This step's noted accuracy was insufficient compared with the recently published techniques; therefore, we optimized the NNs using BO. Also, we increased the hidden layers for each classifier.

E. BO-Based Learning

Black box optimization problems, in which the objective function, represented by the symbol $f(y)$, is treated as a black box, are the focus of deep learning optimization. In such scenarios, BO is shown to be extremely beneficial, especially when human expertise cannot significantly enhance accuracy. By combining previous knowledge of the function f and continuously updating posterior information, this method minimizes loss and maximizes model accuracy. In contrast to the difficult and nonreproducible process of human tuning, BO effectively identifies the global optima of the black box function of the NN. The first step of the BO is a Gaussian process to update the prior function F results and adopt the posterior distribution.

Gaussian processes are deep learning techniques developed using Bayesian learning theory and Gaussian stochastic processes. Any finite subcollection of random variables has a multivariate Gaussian distribution for a stochastic process called a Gaussian process. A statistical model of the function is assumed by the Gaussian process, which is based on the idea that comparable input produces similar output. Similar to a Gaussian distribution defined by mean and covariance, a Gaussian process is defined by its mean function, $n : \mu \to R$, and its covariance function, $p : \mu \times \mu \to R$. The Gaussian process is known as follows:

$$f\left(\mu\right) \sim GP\left(n\left(\mu\right), p\left(\mu, \mu'\right)\right). \tag{13}$$

The probability density function $f(\mu)$ for an arbitrary μ is no longer a scalar but rather a normal distribution function over all possible values of $f(\mu)$. This is how the Gaussian process differs from the Gaussian distribution assume for convenience that the mean function $n(\mu)$ of the Gaussian process is 0.

The exponential square function is a common option for the covariance function p

$$p\left(\mu i : \mu j\right) = \exp\left(-\frac{1}{2}\mu i : \mu j^2\right). \tag{14}$$

The following procedure can be used to determine $f(\mu)'$ posterior distribution. As the training set $Z1$, the first sample t observations are as follows: $Z_{1:t} = \{\mu_n, f_n\}_{n=1}^t, f_n = (\mu_n)$. Assume, a multivariate normal distribution is used to draw the function values f, which are drawn according to a multivariate normal distribution $f \sim M(0, R)$.

$$R = \begin{bmatrix} p\left(\mu_1, \mu_1\right)\left(\mu_1, x_2\right)\ldots.p\left(\mu_1, \mu_t\right) \\ p\left(\mu_2, \mu_1\right)\left(\mu_2, x_2\right)\ldots.p\left(\mu_2, \mu_t\right) \\ \ldots \\ \ldots \\ p\left(\mu_t, \mu_1\right)\left(\mu_t, \mu_2\right)\ldots.p\left(\mu_t, \mu_t\right) \end{bmatrix}. \tag{15}$$

At the new sample point μ_{t+1}, compute the function value $f_{t+1} = f(\mu_{t+1})$ based on the function f_1. The function value f_{t+1} and $f_{1:t}$ in the training set, under the Gaussian process assumption, follow the t+1 dimensional normal distribution

$$\begin{bmatrix} f_{1:t} \\ f_{t+1} \end{bmatrix} \sim M\left(0, \begin{bmatrix} R & p \\ p^T & p\left(\mu_{t+1}, \mu_{t+1}\right) \end{bmatrix}\right) \tag{16}$$

where $f_{1:t} = [f_1, f_2, \ldots \ldots .. f_t]^T$,

$$p = [p\left(\mu_{t+1}, \mu_1\right) p\left(\mu_{t+1}, \mu_2\right)\ldots.p\left(\mu_{t+1}, \mu_t\right)]. \tag{17}$$

And f_{t+1} follows a one-dimensional normal distribution, i.e., $f_{t+1} =\sim M(\eta_{t+1}, \beta^2{}_{t+1})$. By the properties of the joint Gaussian (normal distribution).

$$\eta_{t+1}\left(\mu_{t+1}\right) = p^T R^{-1} f_{1:t} \tag{18}$$

$$\beta^2{}_{t+1}\left(\mu_{t+1}\right) = p^1 R^{-1} p + \left(\mu_{t+1}\mu_{t+1}\right). \tag{19}$$

In the second step, the best points are chosen for the function F by using an acquisition function. We opted for the EI activation function in this work. When a point is explored in the region of the present optimum value, function EI computes the EI it can achieve. The difference between the function value at the sample point value and the present optimum value is known as the degree of improvement EI. The improvement function is 0 if the function value at the sample point value is smaller than the existing optimum value. Mathematically, it is defined as follows:

$$I\left(\mu\right) = \max\left\{0, f_{t+1}\left(\mu\right) - f\left(\mu^+\right)\right\}. \tag{20}$$

Our aim in this function is to maximize EI about the existing optimum value $f(\mu^+)$.

$$\mu = \operatorname{argmax} E\left(\max\left\{0, f_{t+1}\left(\mu\right) - f\left(\mu^+\right)\right\}\right) \tag{21}$$

where $f_{t+1}(\mu) - f(\mu^+) \geq 0$, when the distribution of $f_{t+1}(\mu)$ with mean $\eta(\mu)$ and standard deviation $\beta^2(\mu)$ follows the normal distribution. Therefore, the normal distribution with mean and standard deviation $\beta^2(\mu)$ is the distribution of the random variable EI. EI is defined as follows:

$$E\left(.\right) = \int_\infty^\infty EI.f\left(l\right) dl$$

TABLE II
SELECTED HYPERPARAMETERS AND IT RANGES

Hyperparameters	Ranges
Learning rate	[0.0001 ,1]
Section depth	[0.01, 3]
Momentum	[0.3, 0.98]
L2Regularization	$[1e^{-5}, 1e^{-1}]$
Dropout	[0.001, 0.4]
Activation type	ReLU, Sigmoid, Leaky ReLU

$$= \int_{l=0}^\infty l\frac{1}{\sqrt{2\pi\beta\left(\mu\right)}}exp\left(\frac{\eta\left(\mu\right) - f\left(\mu^+\right) - \text{EI}^2}{2\beta^2\left(\mu\right)}\right) \tag{22}$$

$$dl = \beta\left(\mu\right)\left[Y\emptyset\left(Y\right) + \emptyset\left(Y\right)\right] \tag{23}$$

$$Y = \frac{\eta\left(\mu\right) - f\left(\mu^+\right)}{\beta\left(\mu\right)}. \tag{24}$$

In the third step, the suggested area for sampling produced by the acquisition function is determined. In the fourth step, use an objective function to validate the results. We are finally adding the previously chosen data to the best optimized sample points and modifying the statistical Gaussian distribution model. In this work, we used BO for the dynamic selection of hyperparameters for the training of the proposed model. The selected hyperparameters are listed in Table II.

After the fine-tuning of selected NN classifiers using BO, classification has performed again and obtained some improved accuracy; however, there is a drawback: computational time. Therefore, we proposed a new optimization technique named QHPO that selects the best testing features of the self-attention layer.

F. QHPO-Bases Feature Selection

The original HPO is a population-based optimization algorithm in which search agents are hippopotamus (HP). The update in the position of each HP represents the values for the decision variables. Hence, each HP denotes a vector, and a matrix defines the population of HP. Similarly, to other optimization algorithms, the random initial solutions have been generated; hence, the vector of the decision variables is generated using the following formulation:

$$\xi_i : \tilde{\phi}_{ij} = L_j + r.\left(uL_j - lL_j\right) \tag{25}$$

where $i = 1, 2, 3, .., N$ and $j = 1, 2, 3, \ldots, M$. The notation ξ_i denotes the position of the ith candidate solution, r denotes the random numbers, uL_j denotes the upper bound of the jth solution decision variable, lL_j denotes the lower bound of the jth decision variable, N denotes the total population size, and m denotes the total number of decision variables in the problem.

$$\xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ . \\ \xi_i \\ \xi_N \end{bmatrix}_{N\times m}. \tag{26}$$

1) Phase 1. Update Position (Exploration): In this phase, the male HP position has been updated by employing the following formulation:

$$\xi_i^{male} : \tilde{\phi}_{ij}^{male} = \tilde{\phi}_{ij} + \tilde{\phi}_1 \cdot \left(DHP - I_1 \tilde{\phi}_{ij} \right) \quad (27)$$

where ξ_i^{male} denotes the male HP position and DHP denotes the dominant HP positions, respectively. After that, the female HP position has been updated as follows:

$$\xi_i^{Fmale} : \tilde{\phi}_{ij}^{Fmale} = \begin{cases} \tilde{\phi}_{ij} + H_1 . \left(DHP - I_2 \mu \right) & T > 0.625 \\ Ignore, & Elsewhere \end{cases} . \quad (28)$$

The ξ_i^{Fmale} denotes the position of female HP, μ is a mean value of a few randomly selected HP, and T is a threshold value that is selected based on the different trails. In the later step, the immature HP position has been updated based on the following objective function:

$$\xi_i = \begin{cases} \xi_i^{male} & F_i^{male} < F_i \\ \xi_i & Elsewhere \end{cases} \quad (29)$$

$$\xi_i = \begin{cases} \xi_i^{Fmale} & F_i^{Fmale} < F_i \\ \xi_i & Elsewhere \end{cases} . \quad (30)$$

2) Phase 2. HP Defense Against Predators: Security and safety of the HP is another factor; therefore, it is important to protect them by heavy-weighted animals. Hence, the predator position is an important factor and it is defined as follows:

$$\widetilde{PR} : \widetilde{PR}_j = lL_j + \vec{r}_8 . \left(uL_j - lL_j \right) \quad (31)$$

where $\vec{r}_8$ denotes a random vector of range 0 to 1 and range of $j = 1, 2, 3, \ldots, m$. The distance is computed after the following mathematical equation:

$$\overrightarrow{Dis} = \left| \widetilde{PR}_j - \tilde{\phi}_{ij} \right| . \quad (32)$$

The distance is computed among ith HP to the predator. The updated HP position is further performed using the following mathematical equation:

$$\xi_i^{HPR} : \tilde{\phi}_{i,j}^{HPR} =$$
$$\begin{cases} \overrightarrow{RV} \oplus \widetilde{PR}_j + \left(\frac{e}{(c - d \times \cos(2\pi g))} \right) \left(\frac{1}{\overrightarrow{Dis}} \right) F_{\widetilde{PR}_j} & < F_i \\ \overrightarrow{RV} \oplus \widetilde{PR}_j + \left(\frac{e}{(c - d \times \cos(2\pi g))} \right) \left(\frac{1}{2 \times \overrightarrow{Dis} + r_9} \right) F_{\widetilde{PR}_j} & \geq F_i \end{cases} \quad (33)$$

where ξ_i^{HPR} is HP position which was focused on the predator, $\overrightarrow{RV}$ denotes the reduction vector based on the Leavy distribution (LD). In the next step, a mitigation process has been conducted as follows:

$$\xi_i = \begin{cases} \xi_i^{HPR} F_i^{HPR} & < F_i \\ \xi_i F_i^{HPR} & \geq F_i \end{cases} . \quad (34)$$

3) Phase 3. Exploitation With Quantum Gate: In this phase, the new positions update formulation has been discussed. The HP which are immature or adult leave the group, they faced by an attack of predator. Therefore, the position is updated as follows:

$$lL_j^{local} = \frac{lL_j}{t}, uL_j^{local} = \frac{uL_j}{t}, \text{where } t = 1, 2, 3, \ldots, T \quad (35)$$

$$\xi_1^{HP\in} : \tilde{\phi}_{i,j}^{HP\in} = \tilde{\phi}_{i,j} + r_{10}$$
$$\cdot \left(lL_j^{local} + \mathbb{S}_1 \left(uL_j^{local} - lL_j^{local} \right) \right) \quad (36)$$

where $\xi_1^{HP\in}$ denotes the position of HP used to find the search space of the nearest safe place. Hence, the final activation function is defined as follows:

$$\xi_i = \begin{cases} \xi_i^{HP\in} F_i^{HP\in} & < F_i \\ \xi_i F_i^{HP\in} & \geq F_i \end{cases} . \quad (37)$$

The selected feature vector $(\xi_i \ F_i^{HP\in})$ is further refined using the rotation quantum gate (RQG) approach. In the quantum modification, we used RQG to speed up HP's search space. Each Q-bit individual balances exploitation and exploration to speed up the search process and maintain a distance from the predator. The RQG is mathematically defined as follows:

$$\mathbb{U}(g) = \begin{bmatrix} Cos(g) & -Sin(g) \\ Sin(g) & Cos(g) \end{bmatrix} . \quad (38)$$

The rotation angle (g) and direction of g is determined from the lookup table [44]. This equation is embedded with (36) and updates the position of selected features (HP). The following fitness function is employed for each binary feature vector as follows:

$$Fitness = \alpha \left((Err) + \beta (k) \right) . \quad (39)$$

And cost functions are defined as follows:

$$\oint_{cost} = \phi_\alpha \times \eta_{error} + \phi_\beta \times \left(\frac{num\_feat}{max\_feat} \right) \quad (40)$$

$$\eta_{error} = 1 - \mathcal{A}_{accuracy} \quad (40a)$$

where ϕ_α and ϕ_β are constant variables and presented the values of ϕ_α is 0.99 and ϕ_β is 0.01, $\oint_{cost}$ denoted the cost function and $\mathcal{A}_{accuracy}$ denoted the obtained accuracy from the fitness function. The features are selected by applying the QHPO and the size of selected features are $N \times M$. The selected features are finally classified using fine-tuned BO hyperparameters selected NN classifiers. The detailed results are presented in Section III.

III. RESULTS AND DISCUSSION

The results of the proposed framework have been described in this section. The experiments are carried out on three selected datasets, as discussed in dataset description Section II-A. Each dataset is separated into 50:50. This indicates that 50% of samples are used for training, and the remaining 50% are utilized for testing. The proposed model is 134 layers deeper with 18.6M trainable parameters. The 10-k fold cross-validation is employed to prevent the overfitting problem. For the training of the proposed model, the hyperparameters are learning rate, mini-batch size, epochs, and optimizer with the values of 0.00012, 128, 40, and SGDM. NN classifiers are selected for the classification

TABLE III
SELECTED HYPERPARAMETERS FOR TUNING USING BO

Name	Range	Type
No. of layers	[1,3]	Integer
Activations	RELU, tanh, Sigmoid	Categorical
Standardize	[true, false]	Categorical
Lambda	[9.5238e-09,95.2381]	Categorical
Layer_1_Size	[1,300]	Real

TABLE IV
CLASSIFICATION RESULTS OF THE PROPOSED FUSED SELF-ATTENTION CNN
MODEL ON THE EUROSAT DATASET

Classifiers	Precision (%)	Recall (%)	F1-score	Accuracy (%)	Time (Sec)
NNN	83.05	85.04	84.03	85.1	1185.9
MNN	87.14	90.14	88.61	88.9	400.67
WNN	**89.05**	**88.59**	**88.82**	**90.3**	**493.57**
BNN	84.54	84.32	84.40	85.6	1090
TNN	83.63	83.61	83.60	85.7	1154.4

Bold values shows the best results.

outcomes. The BO is employed for the hyperparameter tuning of NN classifiers. The selected hyperparameters for tuning are described in Table III. The classification outcomes are evaluated using accuracy, precision rate, Recall rate, F1-score, and computation time in seconds. All the experiments are performed using MATLAB2023b using a Desktop Gigabyte Computer designed with a 13th Gen Core-i5 3.50 GHz processor, 128 GB RAM, 500 GB SSD, 1 TB HDD, and 12 GB NVIDIA RTX 3060 graphic card.

The following experiments have been performed to evaluate the proposed framework.

1) *Experiment 1:* Proposed classification results using proposed fusion-based self-attention CNN architecture without using optimized hyperparameters of NN classifiers.
2) *Experiment 2:* Proposed classification results using fusion-based self-attention CNN architecture using BO-based optimized hyperparameters of NN classifiers.
3) *Experiment 3:* Proposed classification results using QHPO (features selection, whereas the BO-based optimized hyperparameters of NN classifiers are chosen).

A. Classification Results on the EuroSAT Dataset

The classification results of the proposed architecture for the EuroSAT dataset are presented in this section. The results of each of the listed experiments have been explained. In experiment 1, the proposed network-level fusion self-attention CNN was trained on the EuroSAT dataset, and the self-attention features were extracted. Many NNs such as narrow NN (NNN), medium NN (MNN), wide NN (WNN), bilayered NN (BNN), and trilayered NN (TNN) have been employed and computed the classification results. The results of this experiment are given in Table IV. This table describes that the WNN classifier achieved the highest accuracy of 90.3%. The precision rate is 89.05%, the recall rate is 88.59%, and F1-score is 88.82%. These measures are also noted for the rest of the listed classifiers. The

TABLE V
CLASSIFICATION RESULTS OF BO TUNING OF NN CLASSIFIERS ON THE
EUROSAT DATASET

Classifiers	Precision (%)	Recall (%)	F1-score	Accuracy (%)	Time (Sec)
NNN	87.05	87.06	87.05	88.0	1828.3
MNN	87.49	87.46	87.47	88.4	1860.8
WNN	89.06	89.11	89.08	89.9	2278.3
BNN	87.98	87.98	87.98	88.9	2212.6
TNN	**89.01**	**89.14**	**89.07**	**90.0**	**2121.2**

Bold values shows the best results.

Fig. 7. Confusion matrix of EuroSAT dataset after employing QHPO-optimized architecture.

computation time is also recorded for all the NN classifiers. The shortest time is recorded for the MNN classifier at 493.57 (sec), while the longest time is noted for the NNN classifier at 1185.9 (sec).

To improve the accuracy and other performance measure, we fine-tuned the hyperparameters of these selected NN classifiers in Experiment No 2. The hyperparameters have been fine-tuned using a BO algorithm. The improved classification results of NN classifiers are shown in Table V. The table statistics show that the TNN classifier obtained higher accuracy than the other classifiers. The accuracy of TNN is 90.0%, precision is 89.01%, recall is 89.14%, and F1-score is 89.07%. The execution is recorded for all the classifiers, and it is observed that this process increases the time; however, the accuracy and other performance measure values have been improved. The longest time of experiment 1 is 1154.4 (sec); however, experiment 2's longest time is 2278.3 (sec).

To further reduce the computational time of the proposed architecture, we implemented a new QHPO algorithm that selects the best features for the classification. Table VI presents the classification results of the proposed QHPO. From this table, the WNN classifier gained 89.5% accuracy. The precision, recall, and F1-score values are 88.63%, 88.65%, and 88.63%, respectively. Moreover, the confusion matrix of this experiment is also illustrated in Fig. 7. The confusion matrix gives the number of

TABLE VI
CLASSIFICATION RESULTS OF THE PROPOSED QHPO ON THE EUROSAT DATASET

Classifiers	Precision (%)	Recall (%)	F1-score	Accuracy (%)	Time (Sec)
NNN	85.29	85.27	85.27	86.3	284.01
MNN	86.09	86.95	86.51	87.8	193.06
WNN	**88.63**	**88.65**	**88.63**	**89.5**	**190.37**
BNN	84.70	84.81	84.81	85.8	228.18
TNN	85.24	85.25	85.24	86.2	230.52

Bold values shows the best results.

TABLE VII
CLASSIFICATION RESULTS OF THE PROPOSED FUSED CNN ON NWPU-RESISC45 DATASET

Classifiers	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Time (Sec)
NNN	82.9	82.9	82.9	82.3	239.18
MNN	86.8	86.7	86.7	86.1	102.14
WNN	**88.3**	**88.1**	**88.2**	**87.7**	**148.63**
BNN	82.8	82.5	82.6	82.1	371.63
TNN	81.8	81.9	81.8	81.4	360.73

Bold values shows the best results.

TABLE VIII
CLASSIFICATION RESULTS OF BO TUNING OF NN CLASSIFIERS ON NWPU-RESISC45 DATASET

Classifiers	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Time (Sec)
NNN	89.76	89.1	89.42	89.9	163.93
MNN	91.44	91.00	91.21	91.2	148.83
WNN	**91.81**	**91.51**	**91.65**	**91.8**	**186.18**
BNN	89.05	88.45	88.74	89.1	357.61
TNN	88.85	88.37	88.60	89.1	347.09

Bold values shows the best results.

observations and TPR values. The confusion matrix can confirm the TNN classifier's accuracy and other performance measures. The computation time is measured for all the classifiers, and it is observed that the QHPO-based feature selection process significantly reduced the computation time and almost maintained the accuracy. The longest time for this experiment is 284.01 (sec) for the NNN classifier, and the WNN classifier executed in a minimum execution time of 193.06 (sec).

B. Classification Results on NWPU-RESISC45 Dataset

The classification results of the proposed architecture for the NWPU-RESISC45 dataset are presented under this subsection. The results of each of the listed experiments have been explained. In the first experiment, the proposed fused Self-Attention CNN architecture is trained on the NWPU-RESISC45 dataset, and the eminent features are extracted from the self-attention activation. The NN classifiers are utilized to obtain the classification results, as presented in Table VII. From this table, it is observed that the WNN classifier obtained a maximum accuracy of 87.7%. WNN's precision, recall, and F1 scores are 88.3%, 88.1%, and 88.2%, respectively. These values are also measured for all

the listed classifiers. The execution time is recorded, and it is observed that the BNN classifier requires 371.63 (sec) for the execution, and the MNN classifier needs 102.14 (sec) for carry out.

In the second experiment, the BO is utilized to optimize the hyperparameters of NN classifiers. Table VIII shows the optimized results of NN classifiers on the NWPU-RESISC45 dataset. The WNN classifier achieved a higher accuracy of 91.8%. The other parameters, such as precision, recall, and F1-score, are also measured, with 91.81%, 91.51%, and 91.65%, respectively. The accuracy of the WNN classifier improves from 87.7% to 91.8% after fin-tuned NN classifiers using BO. The execution time is noted for all the classifiers, and it was observed that, with the improvement of accuracy, the execution time increased by ∼37 (sec) for the WNN classifier.

To reduce the computation time of the fine-tuned NN classifiers, we implemented a QHPO features selection algorithm. The classification results of the proposed QHPO algorithm are presented in Table IX. From this table, the WNN classifier again achieved the maximum accuracy from all the other classifiers, and it was also better in other parameters, including precision, recall, and F1-score. The values of these parameters are 91.91%, 91.44%, and 91.67%, respectively. The confusion

TABLE IX
CLASSIFICATION RESULTS OF THE PROPOSED QHPO ON NWPU-RESISC45 DATASET

Classifiers	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Time (Sec)
NNN	88.75	88.14	88.44	88.7	130.12
MNN	90.77	92.08	91.42	90.6	44.382
WNN	91.91	91.44	91.67	**91.7**	83.809
BNN	87.85	87.35	87.59	87.9	174.07
TNN	88.66	87.93	88.29	88.6	136.15

Bold values shows the best results.

Fig. 8. Confusion matrix of NWPU-RESISC45 dataset after employing the proposed optimized architecture.

matrix of this experiment has been illustrated in Fig. 8. This figure gives the correct prediction rate for each class, including the number of observations. In addition, the computation time is also noted, and it is observed that the computational time is significantly reduced after the selection algorithm. Moreover, the accuracy is also maintained for the features selection algorithm.

C. Classification Results on SIRI-WHU Dataset

The classification results of the proposed architecture for the SIRI-WHU dataset are presented in this subsection. The results of each of the listed experiments have been explained. In the first experiment, the self-attention layer is utilized for feature extraction and fed to NN classifiers for classification. Table X presents the results of this experiment. The WNN classifier gained a maximum accuracy of 98.2%. The other performance measures, such as precision, recall, and F1-score values, are also computed at 98.1%, 98.2%, and 98.1%, respectively. The testing time is computed for all the classifiers, and the TNN

TABLE X
CLASSIFICATION RESULTS OF THE PROPOSED FUSED CNN ON THE SIRI-WHU DATASET

Classifier	Precision	Recall	F1-score	Accuracy	Time (Sec)
NNN	96.9	96.9	96.9	96.9	71.938
MNN	98.0	98.0	98.0	98.1	43.94
WNN	**98.1**	**98.2**	**98.1**	**98.2**	**79.377**
BNN	97.1	97.1	97.1	97.2	99.222
TNN	96.7	96.7	96.7	96.8	222.77

Bold values shows the best results.

classifier takes a long time of 222.77 (sec). The MNN classifier was executed in a minimum time of 43.94 (sec).

To further improve the performance of classifiers, we fine-tuned the hyperparameters of the classifiers using BO.

In the second experiment, hyperparameters of NN classifiers were fine-tuned, and classification was performed. The results of this experiment are presented in Table XI. This table shows that the WNN classifier obtained a maximum accuracy of 98.2%. The

Fig. 9. Confusion matrix of SIRI-WHU dataset after employing the proposed optimized architecture.

TABLE XI
CLASSIFICATION RESULTS OF BO TUNING OF NN CLASSIFIERS ON SIRI-WHU DATASET

Classifier	Precision	Recall	F1-score	Accuracy	Time (Sec)
NNN	97.35	97.33	97.33	97.3	76.413
MNN	97.90	97.87	97.88	97.9	83.382
WNN	**98.19**	**98.16**	**98.17**	**98.2**	**201.43**
BNN	97.92	97.91	97.91	97.9	146.96
TNN	97.67	97.66	97.66	97.7	161.76

Bold values shows the best results.

TABLE XII
CLASSIFICATION RESULTS OF THE PROPOSED QHPO ON SIRI-WHU DATASET

Classifier	Precision	Recall	F1-score	Accuracy	Time (Sec)
NNN	97.95	97.91	97.92	97.9	27.075
MNN	98.27	98.25	98.25	98.2	43.59
WNN	97.24	97.20	97.21	97.2	40.53
BNN	**98.23**	**98.20**	**98.21**	**98.2**	**28.99**
TNN	97.94	97.91	97.92	97.9	89.96

Bold values shows the best results.

precision, recall, and F1-score values are 988.19%, 98.16%, and 98.17%, respectively. The computational time of this experiment is also noted, and it is observed that the time is increased compared with the time of experiment 1. To resolve this challenge, we employed the proposed QHPO features selection algorithm.

The results of experiment 3 are listed in Table XII. This table shows that the BNN classifier achieved the highest accuracy of 98.2% with a 98.23% precision rate, 98.20% recall rate, and 98.21% of F1-score. The obtained accuracy of this experiment is almost consistent, and the confusion matrix is also illustrated in Fig. 9. Using this figure, we can confirm the proposed accuracy of this experiment and other performance measures. The time of this experiment is substantially improved than the previous experiments. The maximum noted time of this experiment is 89.96 (sec) for the TNN classifier, whereas the minimum noted time is 28.99 (sec) for the BNN classifier. Hence, the proposed architecture performed well after employing the feature selection algorithm.

D. Discussion

In this study, we proposed a novel architecture based on the network level fusion and Quantum HPO algorithm for agriculture land cover and land use classification. We designed two parallel CNN models based on inverted bottleneck and dense blocks and fused them using a depth concatenation layer. After that, we embedded a self-attention layer and extracted deep features. Several NN classifiers have been employed for classification accuracy. An inclusive comparison is conducted based on accuracy and computation time with the state-of-the-art deep learning models, as shown in Figs. 10 and 11. We implemented a few pretrained models, such as AlexNet, VGG19, ResNet50, InceptionV3, and NasNetMobile, and compared their accuracy and time with the proposed architecture on selected datasets. Fig. 10 shows that our proposed model outperforms the other networks by achieving 98.2% accuracy on SIRI-WHU, 89.5% on EuroSat, and 91.7% on the NWPU dataset. While the NasNetMobile is a large and complex network from the deep learning networks, it achieved 94.76% accuracy on SIRI-WHU, 86.03% on EuroSat, and 91.7% on the NWPU dataset.
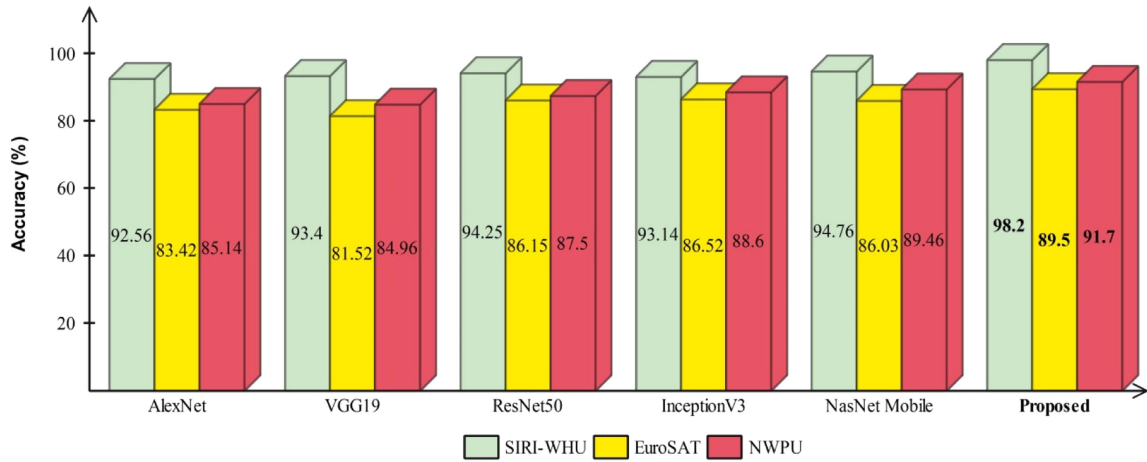
Fig. 10. Comparison of the proposed model accuracy with several other NNs using selected RS datasets.
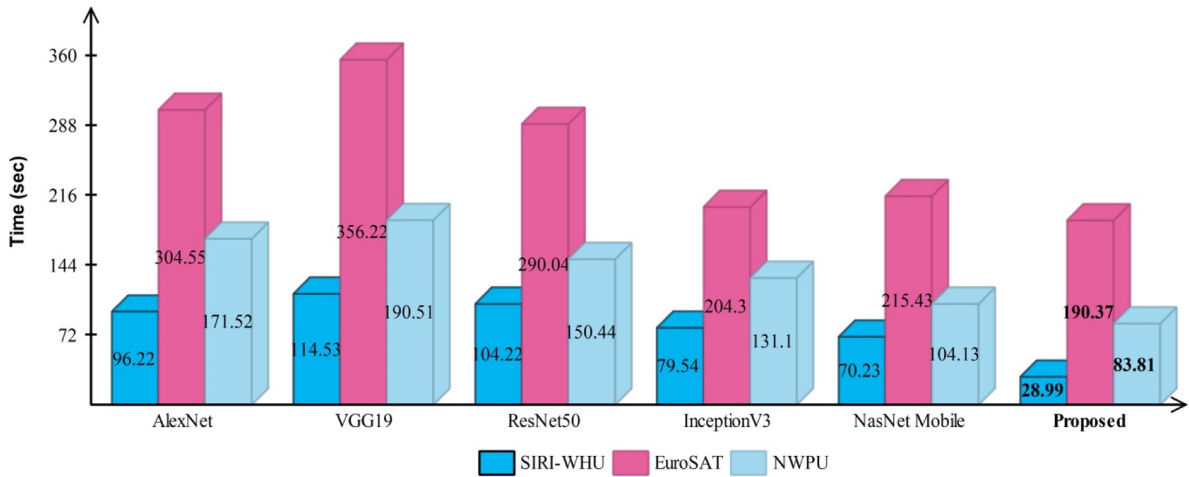

Fig. 11. Comparison of the proposed model testing computational time with several other NNs using selected RS dataset.

The testing computation time is also noted for all the deep learning models. Based on testing computation time, it is seen that our proposed network has taken the shortest time, 190.37 (sec) for EuroSat, 28.99 (sec) for SIRI-WHU, and 83.81 (sec) for the NWPU dataset. From the other deep learning models, VGG19 takes the largest testing time, which is 956.22 (sec) for EuroSAT, 114.53 (sec) for SIRI-WHU, and 190.51 (sec) for the NWPU dataset. Our proposed model is ∼×2 times faster than the state-of-the-art deep learning models. In Fig. 12, It is seen that the proposed model has 134 layers deeper with 18.6M parameters.

We conducted a fair comparison based on datasets with the existing techniques, as presented in Table XIII. According to this table, several techniques are added based on the selected datasets. For the SIRI-WHU dataset, the recent best-attained accuracy was 94.16% by Khan and Basalamah [1]. However, the proposed architecture obtained the best accuracy of 98.20% for this dataset. For the EuroSAT dataset, the recent best accuracy is 88.68% by Zhang et al. [45]. The proposed architecture obtained a maximum accuracy of 89.50%. For the NWPU dataset, Liu et al. [46] obtained a maximum accuracy of 90.30%; however,

TABLE XIII
COMPARISON OF THE PROPOSED ARCHITECTURE ACCURACY WITH
STATE-OF-THE-ART (SOTA) TECHNIQUES

Method	Year	Dataset	Accuracy (%)
[47]	2019	SIRI-WHU	83.88
[48]	2018	SIRI-WHU	92.63
[49]	2016	SIRI-WHU	93.75
[45]	2018	SIRI-WHU	92.08
[50]	2015	SIRI-WHU	91.11
[39]	2023	SIRI-WHU	93.0
[47]	2019	EuroSat	85.23
[48]	2018	EuroSat	87.52
[45]	2018	EuroSat	88.68
[50]	2015	EuroSat	88.51
[51]	2022	SIRI-WHU	88.04
[1]	2023	SIRI-WHU	96.37
[52]	2018	NWPU	82.10
[53]	2019	NWPU	85.70
[46]	2020	NWPU	90.30
Proposed		SIRI-WHU	**98.20**
		EuroSat	**89.50**
		NWPU	**91.70**

Bold values shows the best results.

Models	No. of Layers	Parameters
AlexNet	25	60.9M
VGG-19	47	143.6M
ResNet-50	177	25.5M
InceptionV3	315	23.8M
NasNet Mobile	913	5.3M
Proposed Model	134	18.6M

Fig. 12. Comparison of the proposed model with pretrained on the basis of number of parameters and layers.

the proposed architecture obtained an accuracy of 91.70%. Based on these facts, it is clear that the proposed method obtained a better accuracy and precision rate on the selected datasets.

IV. CONCLUSION

As an illustration, the resolution of the sensor and the weather can impact the quality of the data obtained by RS. Inaccurate managerial decisions can result from low-quality data, which can also impact the accuracy of AI models. In this work, we proposed a novel architecture based on the network-level fused self-attention CNN architecture for agriculture land cover and land use classification. A contrast enhancement technique has been proposed at the initial stage with two important aims to improve the quality of images and to increase the training data. A fused Self-Attention CNN architecture is proposed and trained on the augmented training set. Deep features are extracted from the self-attention layer and classified using NN classifiers. At this stage, we obtained the maximum accuracy of 90.3, 87.7, and 98.2% for EuroSAT, NWPU, andSIRI-WHU datasets. We used BO to improve the NN classifiers' performance further. After BO, the obtained accuracies of the NNs are 90.0, 91.8, and 98.2, and the highest values of precision, recall, and F1-score are 98.1, 98.2, and 98.1, respectively. However, this step increased the computational time; therefore, we proposed a QHPO feature selection algorithm and obtained the best accuracy of 89.5, 91.7, and 98.2% within 200% less computational time with the highest precision and recall rate is 98.23, and 98.20. Overall, the proposed architecture obtained improved accuracy and consumed less time for execution. The limitation of the proposed work is the CNN architecture which required a large amount of labeled data to achieve reasonable outcomes. In the future, a vision transformers-based architecture will be designed for the land cover and land use classification.

DATA AVAILABILITY

The datasets are publically available at Eurostat (https://www.kaggle.com/datasets/apollo2506/eurosat-dataset/), NWPU_R-ESISC45 (https://figshare.com/articles/dataset/NWPU-RES-ISC45_Dataset_with_12_classes/16674166) and the SIRI_WHU (https://figshare.com/articles/dataset/SIRI_WHU_Dataset/8796980).

REFERENCES

[1] S. D. Khan and S. Basalamah, "Multi-branch deep learning framework for land scene classification in satellite imagery," *Remote Sens.*, vol. 15, 2023, Art. no. 3408.

[2] S. Zhao, K. Tu, S. Ye, H. Tang, Y. Hu, and C. Xie, "Land use and land cover classification meets deep learning: A review," *Sensors*, vol. 23, 2023, Art. no. 8966.

[3] N. Guo et al., "HFCC-Net: A dual-branch hybrid framework of CNN and CapsNet for land-use scene classification," *Remote Sens.*, vol. 15, 2023, Art. no. 5044.

[4] M. Shahid et al., "Link-quality based energy-efficient routing protocol for WSN in IoT," *IEEE Trans. Consum. Electron.*, to be published, doi: 10.1109/TCE.2024.3356195.

[5] 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 Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 8500105.

[6] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2021.

[7] N. Guo, M. Jiang, L. Gao, Y. Tang, J. Han, and X. Chen, "CRABR-Net: A contextual relational attention-based recognition network for remote sensing scene objective," *Sensors*, vol. 23, 2023, Art. no. 7514.

[8] K. Manimala, I. G. David, and K. Selvi, "A novel data selection technique using fuzzy C-means clustering to enhance SVM-based power quality classification," *Soft Comput.*, vol. 19, pp. 3123–3144, 2015.

[9] S. Ghodratnama and H. Abrishami Moghaddam, "Content-based image retrieval using feature weighting and C-means clustering in a multi-label classification framework," *Pattern Anal. Appl.*, vol. 24, pp. 1–10, 2021.

[10] H. Wu, B. Liu, W. Su, W. Zhang, and J. Sun, "Hierarchical coding vectors for scene level land-use classification," *Remote Sens.*, vol. 8, 2016, Art. no. 436.

[11] L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognit.*, vol. 71, pp. 158–172, 2017.

[12] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[13] C. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2023, Art. no. 5610426.

[14] T. Akram, B. Laurent, S. R. Naqvi, M. M. Alex, and N. Muhammad, "A deep heterogeneous feature fusion approach for automatic land-use classification," *Inf. Sci.*, vol. 467, pp. 199–218, 2018.

[15] R. Vohra and K. Tiwari, "Land cover classification using multi-fusion based dense transpose convolution in fully convolutional network with feature alignment for remote sensing images," *Earth Sci. Inform.*, vol. 16, pp. 983–1003, 2023.

[16] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinf.*, vol. 19, pp. 1236–1246, 2018.

[17] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. 2nd Int. Workshop Hum. Behav. Understanding*, 2011, pp. 29–39.

[18] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "DeepSat: A learning framework for satellite imagery," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2015, pp. 1–10.

[19] P. Raghavan and N. El Gayar, "Fraud detection using machine learning and deep learning," in *Proc. Int. Conf. Comput. Intell. Knowl. Economy*, 2019, pp. 334–339.

[20] A. Hamza et al., "An integrated parallel inner deep learning models information fusion with Bayesian optimization for land scene classification in satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9888–9903, 2023.

[21] I. Papoutsis, N. I. Bountos, A. Zavras, D. Michail, and C. Tryfonopoulos, "Benchmarking and scaling of deep learning models for land cover image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 250–268, 2023.

[22] Z. Ma, M. Xia, H. Lin, M. Qian, and Y. Zhang, "FENet: Feature enhancement network for land cover classification," *Int. J. Remote Sens.*, vol. 44, pp. 1702–1725, 2023.

[23] U. Patel, M. Pathan, P. Kathiria, and V. Patel, "Crop type classification with hyperspectral images using deep learning: A transfer learning approach," *Model. Earth Syst. Environ.*, vol. 9, pp. 1977–1987, 2023.

[24] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[25] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[26] A. Vali, S. Comai, and M. Matteucci, "Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review," *Remote Sens.*, vol. 12, 2020, Art. no. 2495.

[27] C. Zhang et al., "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, 2019.

[28] H. T. Otal, E. Zavar, S. B. Binder, A. Greer, and M. A. Canbaz, "Harnessing deep learning and satellite imagery for post-buyout land cover mapping," 2024, *arXiv:2401.07500*.

[29] V. Vinaykumar, J. A. Babu, and J. Frnda, "Optimal guidance whale optimization algorithm and hybrid deep learning networks for land use land cover classification," *EURASIP J. Adv. Signal Process.*, vol. 2023, 2023, Art. no. 13.

[30] Z. Lv, P. Zhang, W. Sun, J. A. Benediktsson, J. Li, and W. Wang, "Novel adaptive region spectral-spatial features for land cover classification with high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609412.

[31] Y. Xu et al., "Luojia-HSSR: A high spatial-spectral resolution remote sensing dataset for land-cover classification with a new 3D-HRNet," *Geo-spatial Inf. Sci.*, vol. 26, pp. 289–301, 2023.

[32] I. Ali, Z. Mushtaq, S. Arif, A. D. Algarni, N. F. Soliman, and W. El-Shafai, "Hyperspectral images-based crop classification scheme for agricultural remote sensing," *Comput. Syst. Sci. Eng.*, vol. 46, pp. 303–319, 2023.

[33] A. A. Darem, A. A. Alhashmi, A. M. Almadani, A. K. Alanazi, and G. A. Sutantra, "Development of a map for land use and land cover classification of the Northern Border Region using remote sensing and GIS," *Egyptian J. Remote Sens. Space Sci.*, vol. 26, pp. 341–350, 2023.

[34] S. Bibi et al., "MSRNet: Multiclass skin lesion recognition using additional residual block based fine-tuned deep models information fusion and best feature selection," *Diagnostics*, vol. 13, 2023, Art. no. 3063.

[35] S. Vijh et al., "USMA-BOF: A novel bag-of-features algorithm for classification of infected plant leaf images in precision agriculture," *IEEE Robot. Automat. Mag.*, vol. 30, no. 4, pp. 30–40, Dec. 2023.

[36] S. Rehman et al., "A framework of deep optimal features selection for apple leaf diseases recognition," *Comput. Mater. Continua*, vol. 75, pp. 697–714, 2023.

[37] A. Joshi, B. Pradhan, S. Gite, and S. Chakraborty, "Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review," *Remote Sens.*, vol. 15, 2023, Art. no. 2014.

[38] I. Attri, L. K. Awasthi, T. P. Sharma, and P. Rathee, "A review of deep learning techniques used in agriculture," *Ecological Inform.*, vol. 77, 2023, Art. no. 102217.

[39] S. Thirumaladevi, K. V. Swamy, and M. Sailaja, "Remote sensing image scene classification by transfer learning to augment the accuracy," *Meas.: Sensors*, vol. 25, 2023, Art. no. 100645.

[40] A. Hamza et al., "A novel bottleneck residual and self-attention fusion-assisted architecture for land use recognition in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2995–3009, 2024.

[41] U. Zahra, M. Alhaisoni, A. Alasiry, M. Marzougui, and A. Masood, "An integrated framework of two-stream deep learning models optimal information fusion for fruits disease recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3038–3052, 2023.

[42] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

[43] 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 Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[44] A. Abderrahim, E.-G. Talbi, and M. Khaled, "Hybridization of genetic and quantum algorithm for gene selection and classification of microarray data," *Int. J. Found. Comput. Sci.*, vol. 23, pp. 431–444, 2012.

[45] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[46] P. Liu, G. Gou, X. Shan, D. Tao, and Q. Zhou, "Global optimal structured embedding learning for remote sensing image retrieval," *Sensors*, vol. 20, 2020, Art. no. 291.

[47] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[50] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[51] S. Thirumaladevi, K. V. Swamy, and M. Sailaja, "Multilayer feature fusion using covariance for remote sensing scene classification," *Acta IMEKO*, vol. 11, pp. 8–8, 2022.

[52] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, 2018, Art. no. 1243.

[53] R. Imbriaco, C. Sebastian, E. Bondarev, and P. H. de With, "Aggregated deep local features for remote sensing image retrieval," *Remote Sens.*, vol. 11, 2019, Art. no. 493.

Hussain Mobarak Albarakati received the master's and Ph.D. degrees from the University of Connecticut, USA, in 2016 and 2020, respectively.

He is currently a Senior Professor with the Department of Computer Engineering, College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia, where he teaches courses related to AI and embedded systems. He is a Senior AI Researcher related to remote sensing and medical. He has authored or coauthored more than 50 research articles and also a Reviewer for several reputed journals.

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 of the Department of Computer Science, HITEC University Taxila, Taxila, Pakistan. He has authored or coauthored more than 280 publications which have more than 12 000 citations and an impact factor of more than 850 with h-index 61 and i-index 165. 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.

Dr. Attique Khan is a Reviewer of several reputed journals, such as 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*, *EURASIPE 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 HITEC University, Taxila, Pakistan.

He has authored or coauthored four impact factor papers to date. His major research interests include object detection and recognition, video surveillance, medical, and agriculture using deep learning and machine learning.

Faheem Khan received the master's and Ph.D. degrees in computer science from the University of Malakand, Khyber Pakhtunkhwa, Pakistan, in 2011 and 2017, respectively.

He was an Assistant Professor in Pakistan for four years and supervised many papers and students. Since April 2021, he has been an Assistant Professor with the Department of Computer Engineering, Gachon University, Seongnam, South Korea. His research interests include computer networking, wireless networking, MANET, sensor networking, IoT, artificial intelligence, and AI in healthcare systems.

Naoufel Kraiem received the B.S. degree in computer science from the University of Sfax, Sfax, Tunisia, in 1988, the M.S. degree in computer science from the University of Toulouse, Toulouse, France, in 1990, the Ph.D. degree in computer science from the University of Paris 6, Paris, France, in 1995, and the Habilitation degree in computer science from the University of Sorbonne (Paris 1), Paris, France, in 2011.

He is currently a Professor in Computer Science with the College of Computer Science, King Khalid University, Abha, Saudi Arabia. He has almost 32 years of experience in the academy. His research interests include IT adoption and usage information modeling, software engineering, software product lines, requirement engineering, data science, and CASE tools. His research work has been supported by funding from CNRS, INRIA, Ministry of Research and Technology and Industry (MRT), and by the Commission of the European Communities under the ESPRIT Programmes (BUSINESS CLASS) and UNIDO (National Network of Industrial Information Project).

Dr. Kraiem has been invited to present his research works in many countries in North America, Europe, Africa, and the Middle East. He has been a Member of more than 35 program committees. His research works have been published in reputable international journal's editorial boards (around 147 publications and communications).

Leila Jamel received the engineering degree in computer sciences and the Ph.D. degree in computer sciences and information systems from the National School of Computer Sciences, Tunisia, in 2008.

She is currently an Assistant Professor with the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University (PNU), Riyadh, Saudi Arabia. She is a Researcher with RIADI Laboratory-Tunisia. Her research interests include business process reengineering, process modeling, BPM, data sciences, ML, process mining, e-learning, and software engineering.

Dr. Jamel was the Program Leader of the IS program and the ABET and NCAAA accreditation committees with the CCIS, PNU. She worked as a Head of the Information Systems Security Department with the Premier Ministry in Tunisia. She is a Reviewer of many international journals and conferences. She was a Member of scientific/steering committees of many international conferences.

Latifah Almuqren received the M.Sc. degree (Hons.) in computer science from the University of Coventry, Coventry, U.K., in 2013, and the Ph.D. degree in data mining/machine learning/NLP from Durham University, Durham, U.K., in 2021.

She is currently a Lecturer with the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia.

Roobaea Alroobaea received the bachelor's degree (Hons.) in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, in 2008, and the master's degree in information systems and the Ph.D. degree in computer science from the University of East Anglia, Norwich, U.K., in 2012 and 2016, respectively.

He is currently an Associate Professor with the College of Computers and Information Technology, Taif University, Ta'if, Saudi Arabia. His research interests include human–computer interaction, software engineering, cloud computing, the Internet of Things, artificial intelligence, and machine learning.