ReCXTUnet: A Novel Framework for Remote Sensing Image Classification using TransUnet and XGBoost

Diksha G Kumar[1*], Sangita Chaudhari[1]
[1]Computer Engineering Department, Ramrao Adik Institute of Technology, D. Y. Patil Deemed to be University, Navi Mumbai, India
*Corresponding author email mail: dikshagautam.kumar@gmail.com

(Received on 09 April 2025; In final form on 07 June 2025)

DOI: *https://doi.org/10.58825/jog.2025.19.2.242*

Abstract: Remote sensing image classification (RSIC) is crucial for many environmental and urban applications. RSIC can be difficult due to the high variability and dimensionality in remote sensing image data. This paper presents a novel framework that combines Transformer-based U-Net (TransUNet) and eXtreme Gradient Boosting (XGBoost) for RSIC. TransUNet, known for its powerful feature extraction capabilities, efficiently captures contextual and spatial information from remote sensing images. Additionally, XGBoost improves classification accuracy by efficiently managing high-dimensional data. TransUNet was originally designed for image segmentation tasks, instead of classification. Its architecture is designed to excel at segmenting complex details within images. In our proposed framework, we have adapted TransUNet by adding a classification layer. The fully connected layer of TransUNet serves as the base learner for XGBoost, forming a robust framework for efficient RSIC. This hybrid approach, which combines TransUNet and XGBoost, offers multiple benefits. TransUNet maintains complex details and spatial relationships in images, which improves feature representation. XGBoost provides high predictive accuracy and prevents overfitting with the help of gradient boosting algorithm. This combination tackles challenges in RSIC, such as variations in image quality and noise. We evaluated the proposed approach using high-resolution remote sensing images from the RSI-CB 256 and NWPU-RESISC45 datasets. Our findings show that our framework has outperformed other existing baseline models, attaining an impressive classification accuracy of 91% in RSIC. The experimental results indicate that our approach not only enhances classification accuracy but also remains robust against variations in image quality and noise.

Keywords: Remote Sensing Image classification, Deep learning, TransUnet, ResUnet, SwinUnet, XGBoost

1. Introduction

RSIC, has lately drawn significant research interest by harnessing the capabilities of deep learning methodologies (Yang et al., 2010). Previously, conventional techniques relied on handcrafted features derived from image processing methods like colour histograms, SIFT description (Yang et al., 2008), Gabor-based texture features (Yang et al., 2008), and machine learning models like Support Vector Machine (Gualtieri et al., 1999). In recent times, Convolutional Neural Networks (CNNs) have attained substantial popularity within the realm of classification systems, setting them apart from other classification algorithms. This is attributed to CNN's operational similarity with the neural networks in the human brain. In comparison with statistical classification methods, artificial neural network approaches offer distinct advantages. Latest publications on RSIC have showcased a diverse array of deep learning based models, primarily rooted in deep CNN frameworks. Typically, these newly proposed models employ the convolution operation to gather global features from remote sensing images, stacking multiple convolutional layers to enhance network depth and facilitate ease of training.

CNN significantly enhances the classification accuracy of RSIC. Nonetheless, the rising resolution of remote sensing images brings more detailed categories, leading to challenges like greater intraclass diversity and increased interclass similarity. Commonly used models for image classification, such as DeepLab (Chen et al., 2017), SegNet (Badrinarayanan et al., 2017), and Unet (Ronneberger et al., 2015), are useful. As highlighted by (Li et al., 2021) techniques based on deep learning have proven to be mostly more efficient as compared to simpler classification approaches, like the SVM method by authors (Gualtieri et al., 1999). Furthermore, authors have employed CNNs in their study for water body segmentation, avoiding the use of manually crafted features (Miao et al., 2018). Autoencoders were used to obtain complex feature maps of images of high resolution. Additionally, a technique using Faster RCNN to identify water bodies in satellite images, which, as highlighted by the authors, demonstrated better performance when compared with conventional CNN inspired methodologies (Gharbia et al., 2023).

RSIC is considered conceptually as a combination of the first extraction of features and then their classification. Initially, the model obtains important features, enhancing the subsequent classifier's ability to differentiate among these features more effectively. Then, this classifier takes in the extracted features and efficiently identifies them. Feature extraction stands as a crucial component within an image classification system, directly impacting the classification's overall performance. In image classification tasks, the quality of extracted features significantly affects the classification's accuracy. Until the application of neural networks (NNs) to image classification, prior classifications struggled to fully extract information from image features. Thus, image classification emerged as a key research area within this domain. Theoretically, NNs possess the capability to approximate complex functions and effectively tackle the challenge of image feature extraction. CNNs serve as efficient neural network models for the same. The convolution kernels within CNNs, particularly in the convolutional layers, are responsible for the extraction of features. Utilising convolution, CNNs automatically extract image features, while their hierarchical structure learns high-quality features at each layer.

However, despite CNNs being acknowledged as potent feature extraction mechanisms, the layer for classification within CNNs struggles to comprehensively grasp the information from the extracted features. A single classifier often struggles to handle diverse and complex data features effectively. Ensemble learning addresses these limitations by combining multiple classifiers, each processing distinct hypotheses. This union constructs a stronger hypothesis, resulting in superior predictions.

Studies have demonstrated that ineffective classifiers, when integrated, can produce precise estimations provided sufficient data is available (Kearns et al., 1994). The results highlight the superior learning ability of ensemble learning compared with one classifier, like the improved gradient boosting algorithm, Extreme Gradient Boosting Trees (Chen et al., 2016). A hybrid classification method utilising a CNN and XGBoost model for image classification is presented by the authors, where CNN is used for the extraction of features and XGBoost for effective image classification (Ren et al., 2017). It overcomes the limitations of classical models such as SVM and KNN by taking advantage of the hybridisation of the deep learning architecture and the techniques that deploy an ensemble for the advancement of accuracy. The framework exhibits exceptional performance on benchmark data such as CIFAR-10 and MNIST. The MNIST image database contains 70,000 grayscale images of 10 handwritten digits (from 0 through 9), where every image is of size 28 by 28 pixels, typically used for the classification of digits. The CIFAR-10 database contains 60,000 coloured images, of size 32×32 pixels, distributed into 10 classes, and it is considered a benchmark for image classification. Although the framework offers higher feature utilisation and classification accuracy, it too comes with limitations such as increased computational complexity, the necessity for proper hyperparameter adjustment, and the possibility of overfitting.

SandUnet, derived from Attention U-Net, preserves intricate dune details by retaining uncompressed input signals (Tang et al., 2023). It proceeds with dune-type classification, utilising a fine-tuned MobileNet. This MobileNet combines pretrained knowledge with customisation for adapting to sand dune imagery, autonomously classifying each dune image into six types. Method faces challenges due to spectral similarity between dunes and inter-dune areas, making accurate classification difficult. DRSNet, a customised convolutional neural network for Landsat 8 remote sensing applications, optimised for processing small image patches, is discussed in the paper (Chen et al., 2021). This architecture integrates a unique residual inception channel attention block, merging Inception-ResNet and channel attention for enhanced feature extraction. To prevent representational limitations, pooling layers are swapped with reduction modules. Moreover, the strategic use of up-sampling steps before final pooling layers helps recover information lost during earlier down-sampling phases. DRSNet has various limitations, like high computational demand, limited generalisation, and sensitivity to small patches in remote sensing image classification. To obtain distinct image representations, authors have introduced a method called Aggregated Features from Dual Paths (Shaheed et al., 2023). This approach involves using streamlined convolutional neural networks to create a dual-branch feature extraction method with less parameters. Then, apply a unique feature fusion technique in which they combine bilinear pooling and feature connection concepts. This fusion stage helps to learn discriminative image features effectively. However, faces issues with high computational complexity, large memory requirements, and difficulty in training efficient models on diverse datasets. A CNN model using a specialised block, the residual dense attention block (RDAB), to extract discriminative features from remote sensing scenes (Wang et al., 2022) is demonstrated in the paper. Instance-level vectors are generated from these features, emphasising local information relevant to bag-level labels. A multi-instance pooling strategy guided by channel attention is utilized to emphasize on critical data and filter out non-essential features. Optimization of the network is achieved through the cross-entropy loss, ensuring precise output predictions. Gradient disappearance in deep networks and potential overfitting on complex remote sensing datasets often add limitations.

CNNs, while being effective in image processing, sometimes are unable to capture very fine details or complex structures within the images. Due to their dependence on convolutional layers, CNNs focus more on localised patterns, which might limit them from understanding deeper contextual relations or finer texture variations. Therefore, intricate image representations, for instance, fine-grained object information or abstract spatial configurations, could not be well represented, causing misclassification or loss of significant visual details. To counter the above issues, we propose an innovative RSIC approach that integrates TransUnet and XGBoost.

By leveraging the Transformer-based architecture of TransUNet, the proposed framework excels in capturing and preserving fine-grained details and spatial relationships in remote sensing images. This results in improved feature representation, enabling more accurate classification. The TransUNet-XGBoost framework demonstrates robustness to variations in image quality and noise. The novel modification of TransUNet, originally designed for image segmentation, to include a classification layer for RSIC is a significant advancement. This adaptation ensures that the framework is optimised for the unique challenges of classification tasks. The key innovation in this approach is the fusion of a segmentation framework, TransUnet, for the extraction of features and XGBoost for efficient image classification. Utilising the fully connected layer of TransUNet as a base learner for XGBoost creates a streamlined and efficient classification framework. The ensemble effect created by combining TransUnet and XGBoost leads to efficient classification as compared to leveraging either model in isolation. The framework is designed to be robust and generalise well to different remote sensing scenarios.

2. Dataset

For our experimentation, we have utilised two widely recognised image datasets for RSIC: NWPU-RESISC (Cheng et al., 2017) and RSI-CB256 (Li et al., 2017). Figures 1 and 2 illustrate a few samples from the image dataset of NWPU and RSI datasets, respectively. Below is a summary of the main features of both datasets.

The dataset NWPU-RESISC45 established by Northwestern Polytechnical University is an extensively recognised benchmark for RSIC (Cheng et al., 2017). It features 31,500 images organised into 45 distinct categories, with each category containing 700 images. This dataset serves as an essential resource for the research community to design and evaluate diverse data-driven algorithms. Captured in RGB colour space and with dimensions of 256 x 256 pixels, the dataset has several notable characteristics:

1. Its extensive coverage is demonstrated by the multitude of classes and the substantial overall count of images.
2. It captures a large range of variations, including illumination, spatial resolution, viewpoint, object pose, translation, and background occlusion.
3. Each class shows significant internal diversity while maintaining clear similarities with other classes.

The proposed framework is also evaluated on RSICB256 (Li et al., 2017), it features a broad range of images. The dataset has images from four distinct classes, collected from Google Maps. The visuals highlight the dataset's complexity by displaying Points of interest across multiple countries were analysed using remote sensing imagery sourced from both Bing Maps and Google Earth, featuring spatial resolutions which range between 0.22 to 3 meters. This dataset uses a graded classification system. Influenced by the Chinese land-use framework, it is modified to address various object classification needs while functioning as a robust benchmark for research and experimentation. Our method was evaluated on 4,000 labelled and georeferenced images distributed among four classes: dense residential, bridge, airplane and beach.

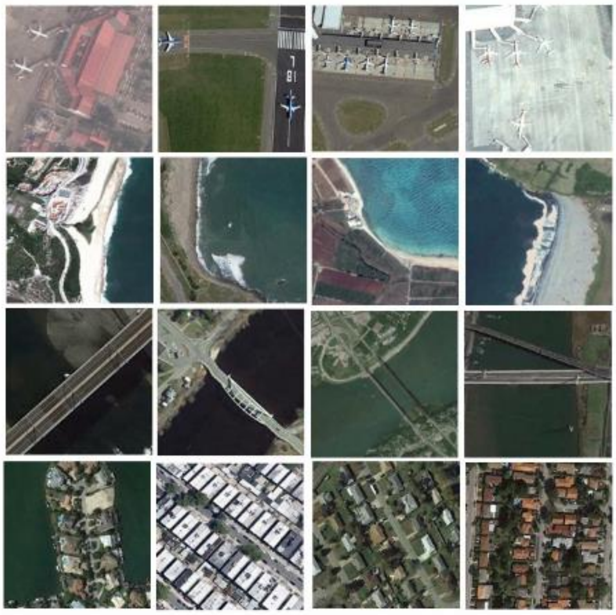


Figure 1: NWPU-RESISC Dataset images: Airplanes, Beaches, Bridges, Dense Residential Areas

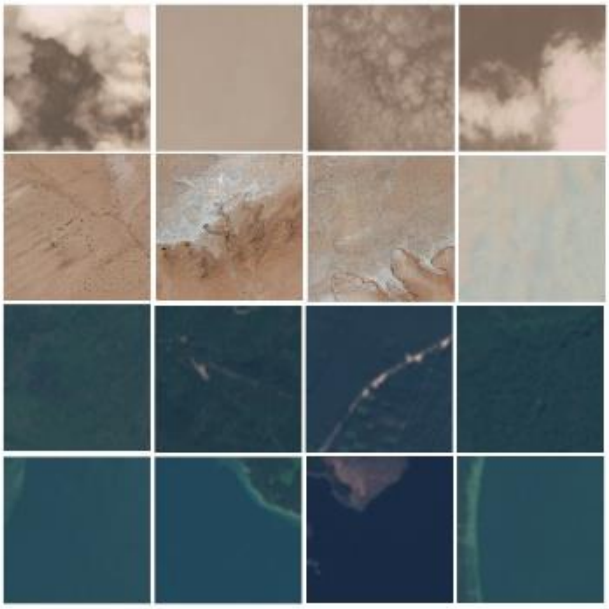


Figure 2: RSI-CB256 Dataset: Top to Bottom - Clouds, Deserts, Green Area, and Water

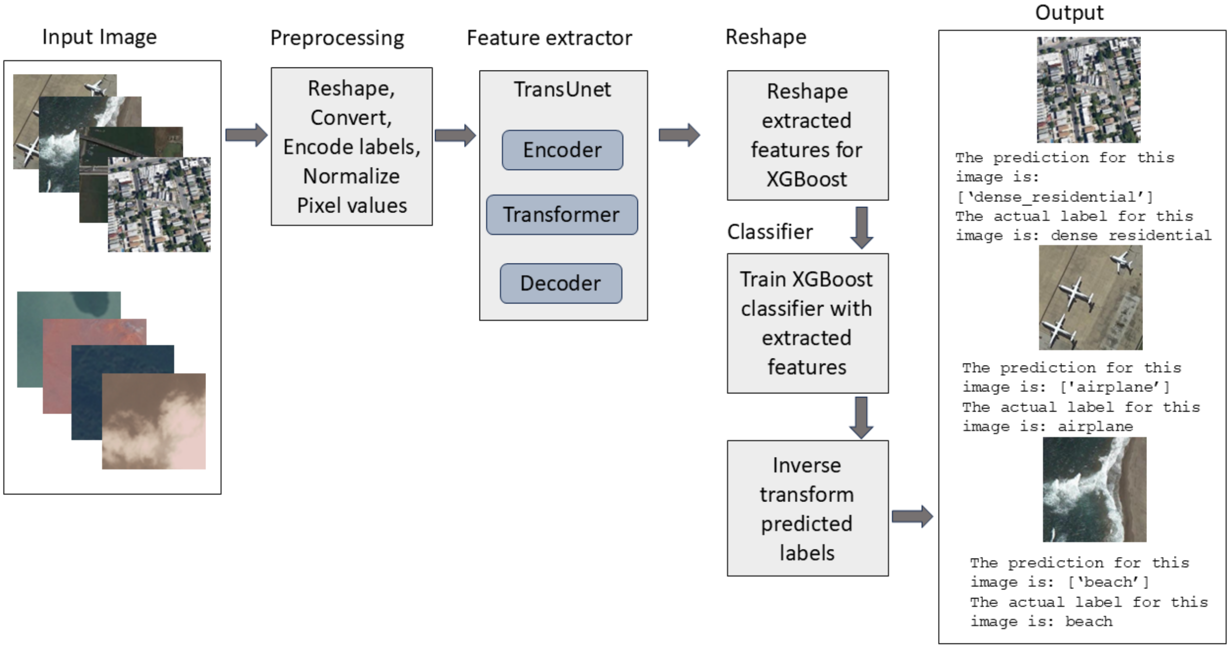

Input Image
Preprocessing
Reshape, Convert, Encode labels, Normalize Pixel values
Feature extractor
TransUnet
Encoder
Transformer
Decoder
Reshape
Reshape extracted features for XGBoost
Classifier
Train XGBoost classifier with extracted features
Inverse transform predicted labels
Output
The prediction for this image is: ['dense_residential'] The actual label for this image is: dense residential
The prediction for this image is: ['airplane'] The actual label for this image is: airplane
The prediction for this image is: ['beach'] The actual label for this image is: beach

Figure 3: The proposed model's systematic flow diagram

3. Materials and methods

The proposed architecture presents a unique method for RSIC. The suggested methodology comprises two primary phases: feature extraction employing TransUnet (Chen et al., 2021) and subsequent classification through XGBoost (Chen et al., 2016). TransUnet is utilised for extracting features, as illustrated in Figure 3. The figure represents the architecture of the proposed framework. Preprocessing is first applied to the image dataset. Normalisation, resizing images to size 256x256, changing images to the RGB colour space, and categorical labels for images are represented in the preprocessing steps. Images are then fed into the initial stage of TransUNet, where the significant features are extracted.

Normalisation during image pre-processing converts pixel intensity values to a standard range, improving contrast and making images suitable for further processing. Normalisation provides consistency in images and thus improves the performance of the model on classification and object detection tasks. The extracted features are then reshaped into a feature vector and input to the classification layer for processing using XGBoost. The inverse transformation is then applied to the labels. Remote sensing images used in this paper were obtained from the RSI-CB256 (Li et al., 2017) and NWPU-RESISC45 (Cheng et al., 2017) datasets. The datasets hold a very diverse set of images ranging over various land cover classes.

The main feature extractor is TransUNet (Chen et al., 2021), which was initially created for image segmentation. It uses remote sensing images to extract contextual and spatial information by utilising the transformer architecture. TransUNet (Chen et al., 2021) is the first medical image segmentation framework that uses sequence-to-sequence prediction to incorporate self-attention mechanisms. By using a hybrid CNN-Transformer design, this framework overcomes the prevalent problem of resolution degradation in Transformer models. It makes use of both the broad contextual insights provided by Transformers and the fine-grained spatial detail captured by Convolutional Neural Networks (CNNs). It enhances the self-attention features produced by Transformers and combines them with high-resolution CNN outputs through skip connections from the encoder pathway, taking inspiration from the U-shaped architecture. Because of this integration, precise localisation is made possible, effectively maintaining the benefits that come with Transformers.

The TransUNet is adapted by incorporating a classification layer into its framework in our paper. TransUNet's fully connected layer is modified to predict class probabilities corresponding to the remote sensing images. The adapted TransUNet processes input images to extract high-dimensional feature representations. These features capture intricate details and spatial relationships within the images. The fully connected layer of the adapted TransUNet acts as the base learner for XGBoost, with the extracted features being input into the XGBoost algorithm for further processing and classification.

XGBoost, an ensemble learning technique built on gradient boosting, is employed for enhancing classification accuracy. It efficiently handles high-dimensional data and provides robustness against overfitting. The extracted features from TransUNet are utilised in the training of the XGBoost classifier. Below are the mathematical representations of each stage, from preprocessing to label prediction.

1. I represent the input remote sensing image. The pre

processing step involves:

$$I' = \text{Preprocess}(I) \qquad (1)$$

Here, I' represents the pre-processed image, with preprocess functioning as the operation that executes the required preprocessing steps.

2. The encoder part of TransUNet, which typically comprises convolutional layers along with a series of transformer blocks

$$Z = f_{\text{encoder}}(I') \qquad (2)$$

where Z denotes the features derived from the image. These features are then processed through the transformer blocks to extract spatial and contextual information effectively.

3. Next, the decoder processes these features

$$S = f_{\text{decoder}}(Z) \qquad (3)$$

where S denotes the output segmentation map, where each pixel in the image is labelled with its corresponding class.

4. After obtaining the segmentation map or features from TransUNet, these extracted features are input to XGBoost for performing the classification task. XGBoost works by learning a decision tree ensemble to predict class labels.

5. F represents a matrix of extracted features.$\hat{Y}$ represents predicted class labels from the XGBoost classifier.

$$\hat{Y} = \text{XGBoost}(F) \qquad (4)$$

where XGBoost(F) denotes the classification process, using the features F for training.

XGBoost (Chen et al., 2016) utilises an ensemble learning methodology that builds an effective predictive model by integrating the predictions of various weaker models. XGBoost is constructed on a gradient boosting framework. It uses decision tree ensembles as its base learners. The decision trees are trained consecutively, where each tree addresses the errors of its predecessor. XGBoost integrates L1 and L2 regularisation terms within its framework to enhance performance and prevent overfitting. XGBoost provides a feature importance score, highlighting the significance of each feature in the model's predictions. This is valuable for feature selection and understanding the data. Cross-validation is often used with XGBoost to assess model performance and select hyperparameters effectively.

The objective of XGBoost is to minimise an objective function that represents the total of the regularisation term and loss function. The XGBoost mathematical function is:

$$Obj(\theta) = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) \;+\; \sum_{k=1}^{k} \Omega(f_k) \qquad (5)$$

Obj(θ) denotes the objective function to be minimised, wherein θ symbolizes a collection of model parameters. f_k are the individual decision trees within the ensemble, while K represents the total count of trees in the ensemble. y_i signifies the true (actual) target label, and $\hat{y}_i$denotes the predicted target label. The function $l(y_i, \hat{y}_i)$ calculates the differences between the predicted and actual values.

4. Results and discussion

The proposed system leverages a hybrid TransUnet and XGBoost model, which captures high-resolution spatial information and global contextual information. This combination is better performing in feature representation and preserving fine-grained detail. The other baseline models employ convolutional layers solely for feature extraction. Even if successful, they cannot capture the global context as efficiently as the Transformer-based models like Trans UNet. TransUnet employs self-attention mechanisms, enabling the model to focus on important regions of the image. This further boosts the accuracy of the classification by focusing on important features. In comparison, other baseline architectures lack inherent attention mechanisms, which may lead to less emphasis on features and lower classification accuracy. XGBoost, an efficient ensemble learning algorithm, further improves the accuracy of classification by handling high-dimensional data efficiently.

For classification accuracy evaluation of TransUnet-XGBoost model and other baseline frameworks, we have implemented some additional baseline models such as Unet (Ronneberger et al., 2015), SwinUnet (Cao et al., 2022) and ResidualUnet (Diakogiannis et al., 2020), which are employed for feature extraction and XGBoost is used for classification. Tables 1 and 2 present the classification accuracies obtained from these baseline models on the utilised datasets, respectively, along with the proposed classification model.

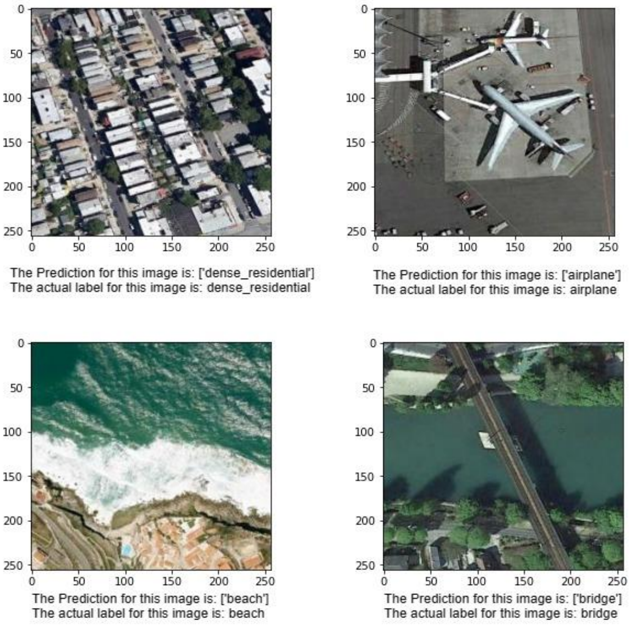
Table 1. Comparison of the proposed framework's accuracy with baseline models on the NWPU dataset

Feature Extraction Method	**Classifier Used**	**Performance Accuracy (%)**
UNet	XGBoost	86.11
ResUNet	XGBoost	82.97
SwinUnet	XGBoost	90.16
TransUNet: Proposed Framework	XGBoost	91.12

Table 2. Comparison of the proposed framework's accuracy with baseline models on the RSI dataset

Feature Extraction Method	**Classifier Used**	**Performance Accuracy (%)**
Unet	XGBoost	86.21
ResUNet	XGBoost	88.50
SwinUnet	XGBoost	85.50
TransUNet: Proposed Framework	XGBoost	89.32

Our proposed framework attained an average classification accuracy of 91.12 % on NWPURESISC and 89.32 % accuracy on the RSI-CB256 datasets, surpassing the other baseline models. The model showcased its ability to classify remote sensing images, demonstrating an impressive ability to correctly classify even when provided with poorly illuminated or visually similar images. The classification prediction results for the four classes across the datasets are illustrated in Figure 4 and Figure 5.



The Prediction for this image is: ['dense_residential']
The actual label for this image is: dense_residential

The Prediction for this image is: ['airplane']
The actual label for this image is: airplane

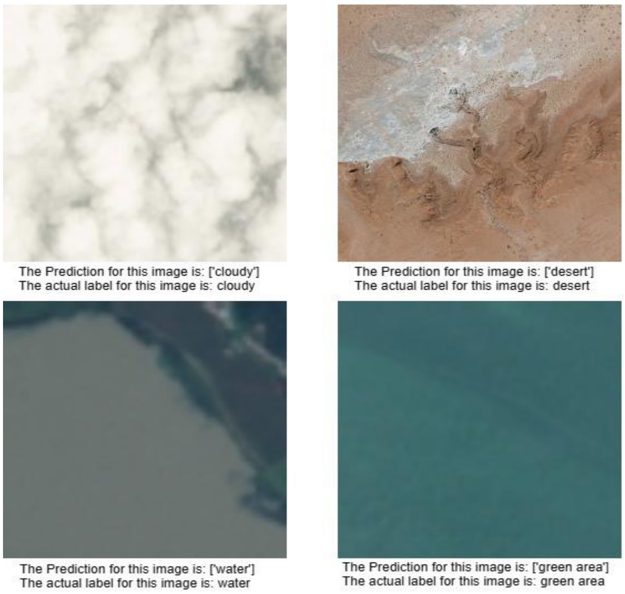
The Prediction for this image is: ['beach']
The actual label for this image is: beach

The Prediction for this image is: ['bridge']
The actual label for this image is: bridge

Figure 4: Prediction results for four classes using NWPU image dataset

In this study, deep learning experimentations are performed utilising the Keras. To keep reliability and fairness, identical hyperparameters were applied across all networks. The TransUNet implementation employed Python 3.7 and TensorFlow 2.2, with the ReLU activation function integrated. Experimental outcomes signify that our framework, TransUnet and XGBoost, generates transferable features with a high degree of robustness, resulting in improved classification accuracy. XGBoost achieves better classification performance due to its high predictability, which can effectively identify the complex relations and characteristics involved in the remote sensing images. At the same time, the proposed model excels in its ability to distinguish images with similar characteristics but having different classes. Robustness of the framework was evaluated using essential evaluation metrics, i.e., recall, F1 score, and precision, as described (Vapnik et al., 1999). They were calculated taking into account a per-pixel confusion matrix for each patch of an image.

Moreover, Tables 3 and 4 showcase the classification reports of all four classes utilised in our proposed system, respectively. These reports provide metrics like Precision, Recall, and F1 scores. Significantly, the dense residential category has attained maximum Precision, F1 score and Recall, on the NWPU and RSI datasets.



The Prediction for this image is: ['cloudy']
The actual label for this image is: cloudy

The Prediction for this image is: ['desert']
The actual label for this image is: desert

The Prediction for this image is: ['water']
The actual label for this image is: water

The Prediction for this image is: ['green area']
The actual label for this image is: green area

Figure 5: Prediction results for four classes using RSI image dataset

Table 3. Performance metrics of the proposed system: Precision, F1 scores and Recall, and of four classes using the NWPU-RESISC45 dataset

Class	Precision	F1 Score	Recall
Airplane	89	91	91
Bridge	89	88	88
Beach	90	88	85
Dense Residential	93	94	96

Table 4. Performance metrics of the proposed system: Precision, F1 scores and Recall, and of four classes using the RSI-CB256 dataset

Class	Precision	F1 Score	Recall
Cloudy	72	83	98
Green Area	98	95	97
Desert	99	76	62
Water	93	95	99

The ROC curve is commonly used methodology for visually assessing classifier performance. Generally, a higher area under the curve of ROC curve reflects better performance, as highlighted by authors (Fawcett et al., 2006). The proposed framework surpasses other baseline models in AUROC for multiple classes. Notably, it attains the highest score, with the Swin-Unet model following at AUROC = 0.94, and ResUNet registering the minimum AUROC value. Interestingly, the proposed framework secures the highest AUROC for the dense residential class. These findings highlight the effectiveness of the proposed framework compared to baseline models. Figure 6 illustrates the AUC scores across all classes for both the proposed framework and the other baseline models, whereas Table 5 summarizes the corresponding AUROC values.

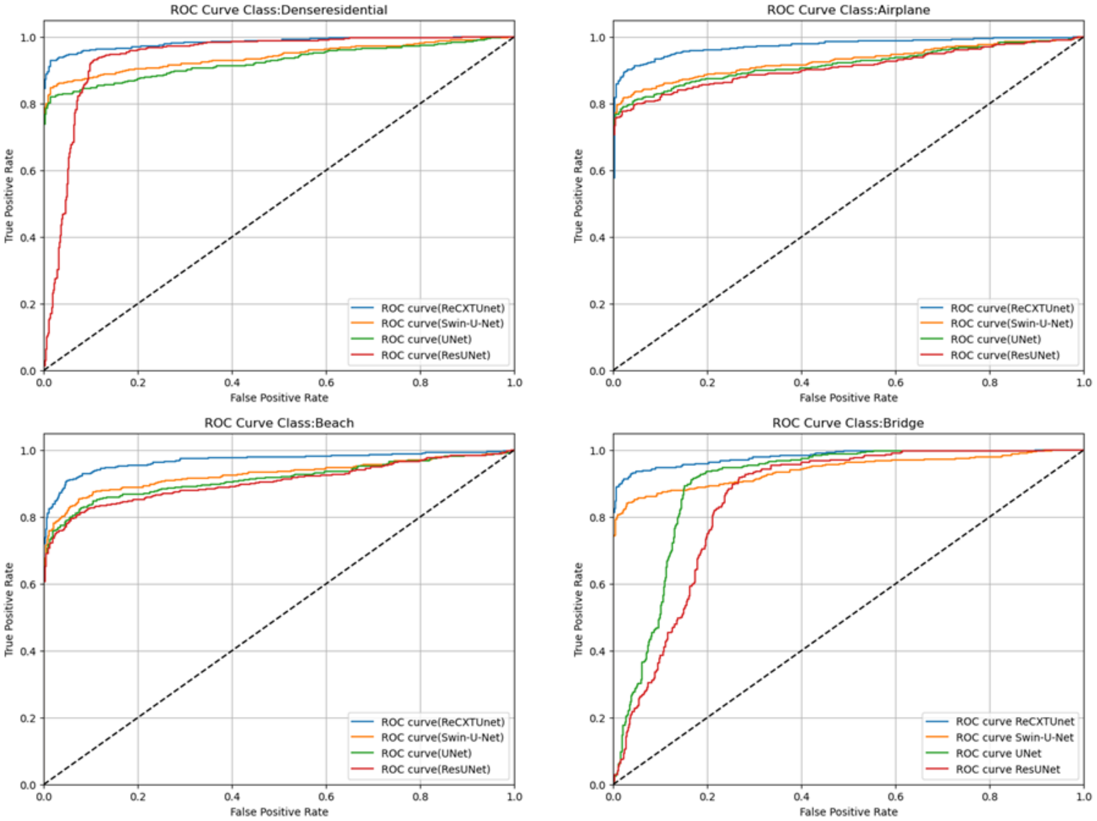


Figure 6. AUROC for the Classes: (a) Dense Residential, (b) Airplane, (c) Beach, and (d) Bridge, Comparing the Proposed Method (ReCXTUnet) with Baseline Models.

Table 5. AUROC performance metrics of the proposed framework versus benchmark models

Class	ResUnet	U-Net	Swin-Unet	TransUNet
Bridge	0.83	0.89	0.93	0.96
Airplane	0.90	0.92	0.94	0.97
Beach	0.91	0.92	0.92	0.97
Dense residential	0.94	0.93	0.94	0.98

In remote sensing, RSIC plays a crucial role across multiple domains, including land and forest analysis, object recognition, urban planning, land use assessment, geological disaster monitoring, map updating, land cover mapping, and natural resource management. RSIC using TransUNet and XGBoost can be effectively applied to urban land cover classification from satellite images. Accurate identification of different land types, like buildings, vegetation, roads, and water bodies, is crucial for urban planning, environmental monitoring, and disaster management. This approach enhances the reliability of urban land cover classification, making it valuable for applications in city development, climate change analysis, and post-disaster damage assessment. Another compelling use case for remote sensing image classification is disaster impact assessment. Following natural calamities like floods, wildfires, storms, and earthquakes, satellite and drone imagery are used to evaluate the extent of damage in affected areas. This approach helps governments and relief organisations prioritise resource allocation, emergency response, and recovery efforts, ultimately supporting faster rehabilitation of disaster-affected communities.

5. Conclusion

Our paper validates the effectiveness of using TransUNet for the extraction of features and the XGBoost classifier for RSIC. TransUNet proved to be a powerful feature extractor, capturing intricate patterns and representations from remote-sensing images. Its ability to learn comprehensive, high-resolution spatial details combined with broad global contextual insights allowed for better discrimination between different land cover categories, enhancing classification performance. XGBoost, as the classification model, showcased its robustness and interpretability, effectively handling the extracted features from TransUNet. The results highlight that this approach outperforms other benchmark models, attaining a classification accuracy of 91 % in classification. The combined TransUNet-XGBoost approach outperformed traditional methods, achieving high classification accuracy.

However, the proposed system has a few limitations. Deep neural networks, including TransUNet, are liable to overfit when trained on limited datasets. Additionally, TransUNet can be seen as a “black box” model, making it challenging to interpret why a particular classification decision was made. Although these limitations exist, our proposed method has demonstrated encouraging results in the classification. Future studies will investigate its potential for broader classification applications, including disaster response, land usage analysis, land coverage classification, and monitoring of the environment.

References

Badrinarayanan, V. and A. Kendall (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481–2495.

Cao, H., Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian and M. Wang (2022). Swin-Unet: Unet-like pure transformer for medical image segmentation. European Conference on Computer Vision, 205–218. Springer.

Chen, F. and J. Y. Tsou (2021). DRSNet: Novel architecture for small patch and low-resolution remote sensing image scene classification. International Journal of Applied Earth Observation and Geoinformation, 104, 102577.

Chen, J., Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille and Y. Zhou (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, P., G. Kok Lim, K. I. Man, M. Khairuddin and Y. L. Chen (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848.

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Cheng, G., J. Han and X. Lu (2017). Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 105(10), 1865–1883.

Diakogiannis, F. I., F. Waldner, P. Caccetta and C. Wu (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162, pp. 94–114.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), pp. 861–874.

Gharbia, R. (2023). Deep learning for automatic extraction of water bodies using satellite imagery. Journal of the Indian Society of Remote Sensing, 51, 1511–1521.

Gualtieri, J. A. and R. F. Cromp (1999). Support vector machines for hyperspectral remote sensing classification. 27th AIPR Workshop: Advances in Computer-Assisted Recognition, 3584, pp. 221–232. SPIE.

Kearns, M. and L. Valiant (1994). Cryptographic limitations on learning boolean formulae and finite automata. Journal of the ACM, 41(1), pp. 67–95.

Li, H., X. Dou, C. Tao, Z. Hou, J. Chen, J. Peng, M. Deng and L. Zhao (2017). RSICB: A large-scale remote sensing image classification benchmark via crowdsource data. arXiv preprint arXiv:1705.10450.

Li, Y., D. Kong, Y. Zhang, R. Chen and J. Chen (2021). Representation learning of remote sensing knowledge graph for zero-shot remote sensing image scene classification. IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, pp. 1351–1354.

Miao, F., K. Sun, H. Song, S. Xu, Y. Ma and Z. Miao (2018). Automatic water-body segmentation from high-resolution satellite images via deep networks. IEEE Geoscience and Remote Sensing Letters, 15(4), pp. 602–606.

Ren, X., H. Guo, S. Li, S. Wang and J. Li (2017). A novel image classification method with CNN-XGBoost model. in Digital Forensics and Watermarking: 16th International Workshop (IWDW 2017), Magdeburg, Germany, 23–25 August 2017, Proceedings 16, pp. 378–390. Springer.

Ronneberger, O., P. Fischer and T. Brox (2015). U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28,

Shaheed, K., Q. Abbas, A. Hussain and I. Qureshi (2023). Optimized Xception learning model and XGBoost classifier for detection of multiclass chest disease from X-ray images. Diagnostics, 13(15), 2583.

Tang, Y., Z. Wang, Y. Jiang, T. Zhang and W. Yang (2023). An auto-detection and classification algorithm for identification of sand dunes based on remote sensing images. International Journal of Applied Earth Observation and Geoinformation, 125, 103592.

Vapnik, V. (1999). The Nature of Statistical Learning Theory. Springer, New York.

Wang, X., H. Xu, L. Yuan, W. Dai and X. Wen (2022). A remote-sensing scene-image classification method based on deep multiple-instance learning with a residual dense attention convent. Remote Sensing, 14(20), 5095.

Yang, Y. and S. Newsam (2010). Bag-of-visual-words and spatial extensions for land-use classification. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems.

Yang, Y. and S. Newsam (2008). Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery. 15th IEEE International Conference on Image Processing, 15(7), pp. 1852–1855.