


# Improvement of Satellite Image Classification Using Attention-Based Vision Transformer

Nawel Slimani<sup>1</sup> <sup>a</sup>, Imen Jdey<sup>2</sup> <sup>b</sup> and Monji Kerallah<sup>3</sup> <sup>c</sup>

<sup>1</sup>National School of Electronics and Telecommunications, Sfax University, Sfax, Tunisia

<sup>2</sup>University of Sfax, ReGIM-Lab. REsearch Groups in Intelligent Machines (LR11ES48), Sfax, Tunisia

<sup>3</sup>Faculty of Sciences of Sfax, Sfax University, Tunisia

**Keywords:** Deep Learning, Classification, Remote Sensing, Computer Vision, Vision Transformer, Self-Attention Mechanism, Satellite Image.

**Abstract:** This study introduces a transformative approach to satellite image classification using the Vision Transformer (ViT) model, a revolutionary deep learning method. Unlike conventional methods, ViT divides images into patches and employs self-attention mechanisms to capture intricate spatial dependencies, enabling the discernment of nuanced patterns at the patch level. This key innovation results in remarkable classification accuracy, surpassing 98% for SAT4 and SAT6 datasets. The study's findings hold substantial promise for diverse applications, including urban planning, agriculture, disaster response, and environmental conservation. By providing a nuanced understanding of ViT's impact on satellite imagery analysis, this work not only contributes insights into ViT's architecture and training process but also establishes a robust foundation for advancing the field and promoting sustainable resource management through informed decision-making.


## 1 INTRODUCTION


The classification of hyperspectral images (HSI) is a very active area of research in remote sensing and earth observation. HSI is distinguished by its extremely high spectral dimensionality (Hang et al., 2020). Satellite imagery can be defined as a representation of a complete part of the earth acquired with artificial satellites (Devi et al., 2023), and it is an important tool for understanding our planet and managing its resources (Scheibenreif et al., 2022) and (Dimitrovski et al., 2023). It provides a unique perspective on the Earth's surface, and it can be used to monitor changes over time and across large areas (Baig et al., 2022). It can offer insightful information for a variety of applications, including emergency management (Daud et al., 2022), land use, and cover analysis (Baig et al., 2022), which provide valuable information about the land surface, including vegetation, water bodies, and urban areas. Image classification helps to identify and map these features, which is important for monitoring changes in land use and land


cover over time. Environmental monitoring can also be used to monitor environmental conditions such as water quality, soil moisture, and air pollution. Satellite image classification helps identify areas where these conditions are changing, which is important for managing natural resources and protecting the environment. Yet, there have not been many attempts to apply visual attention techniques to the study of remote sensing data in the literature (Mehmood et al., 2022).

Deep learning methods, in particular Convolutional Neural Networks (CNNs), have recently demonstrated considerable potential for reaching high accuracy in this job (Bazi et al., 2021) (Slimani et al., 2023). However, these models might fail to capture the image's long-range dependencies, which could lead to incorrect classification or a lack of comprehension of complicated visual structures.

Deep learning's attention mechanism is a potent technique that enables the model to concentrate on particular areas of the input image while ignoring irrelevant data (Zu et al., 2023). The application of this approach to problems involving image identification and natural language processing has been successful. Recently, it has also been applied to satellite image classification to improve deep learning model

<sup>a</sup>  <https://orcid.org/0009-0008-7971-1214>

<sup>b</sup>  <https://orcid.org/0000-0001-7937-941X>

<sup>c</sup>  <https://orcid.org/0000-0003-2451-1721>

performance (Zu et al., 2023). It works by assigning a weight to each element of the input sequence based on its relevance to the output. The elements with higher weights are given more attention, while those with lower weights are given less attention.

The ViT learns to generate a weight vector for each input patch in each self-attention layer of the transformer encoder, depending on how similar it is to every other patch in the sequence. This is accomplished by creating a dot product between each patch's query, key, and value vectors, and then using the softmax function to create a normalized weight vector (Xian et al., 2022).

The weight vector shows how significant each patch is in determining how the feature representation of the image is computed. In contrast to irrelevant or noisy patches, which are given lower weights, patches that are highly relevant to the classification job are given greater weights (Zu et al., 2023). The attended feature representation is then created by computing the input patches' weighted sum using the weight vector as the weights. This enables the model to ignore unimportant or noisy patches and choose to concentrate on the image's most crucial parts.

The feedforward neural network is then fed the attended feature representation to generate the final classification output. The feedforward network applies non-linear modifications to the attended feature representation, which enables the model to accurately predict and capture complicated relationships between the input patches (Xian et al., 2022).

This research paper begins by offering contextual information. Next, we provide an overview of pertinent literature and explore previous methods that influenced our proposed approach. The paper introduces the proposed approach, which involves using an attention-based vision transformer to improve the classification of satellite images in remote sensing monitoring. We present the achieved results and discuss their significance. Lastly, the article summarizes the key findings and proposes potential avenues for future research.

## 2 BACKGROUND

ViT is a recent breakthrough in the area of computer vision (Jamil et al., 2023) (Jdey et al., 2012a). While transformer-based models have dominated the field of natural language processing since 2017, CNN-based models are still demonstrating state-of-the-art performances in vision problems (Li et al., 2022b). Last years, a group of researchers from Google figured out how to make a transformer work on recognition.

They called it a "vision transformer." The follow-up works by the community demonstrated superior performance of vision transformers not only in recognition but also in other downstream tasks such as detection, segmentation, multi-modal learning, and scene text recognition, to mention a few. It is a variant of the popular transformer architecture that was originally developed for natural language processing (NLP) tasks but has since been adapted for computer vision applications such as image classification. According to (Jiang et al., 2019), the main idea behind the ViT is to treat an image as a sequence of patches, each patch being a small, fixed-size subimage of the original image. The patches are flattened into a sequence of 1D vectors, which are then processed by a standard transformer encoder.

The ViT architecture consists of a series of stages, each containing a multi-head self-attention mechanism followed by a feedforward neural network layer. The attention mechanism allows the model to capture long-range dependencies between patches, while the feedforward layer applies non-linear transformations to the representations (Jiang et al., 2019). To improve the quality of the learned features, the ViT model is typically pre-trained on large datasets of images using self-supervised learning techniques. During pre-training, the model learns to predict the relative position of different patches within an image, which encourages the model to learn spatial relationships between patches and improves its ability to recognize objects. Once the model is pre-trained, the final layers can be replaced or fine-tuned for specific downstream tasks, such as image classification or object detection. Overall, the ViT architecture has shown promising results on a variety of image recognition tasks, achieving state-of-the-art performance on several benchmarks while requiring significantly fewer parameters than traditional convolutional neural networks (Jiang et al., 2019).

### 2.1 Attention Mechanism

One common way to use attention in image classification is to apply it to the feature maps produced by the convolutional layers of a deep neural network (Duan and Zhao, 2019). In this approach, the attention mechanism learns to assign weights to different spatial locations in the feature maps based on their relevance to the classification task (Duan and Zhao, 2019).

The attention weights are computed by applying a small neural network to each spatial location in the feature maps. The output of this network is a scalar value, which represents the importance of the corresponding spatial location for the classification task.

The attention weights are then normalized using a softmax function so that they sum to one across all spatial locations.

$$\text{Softmax}(Z)_i = \frac{e^{z_i}}{\sum_{j=1}^N e^{z_j}} \quad (1)$$

Weighted summation of feature maps using attention weights as weights yields a curated feature representation of the image. This curated representation can be used as input to a classifier, such as a fully connected plane, to predict class labels for images. Another way to pay attention to image classification is to apply attention to the intermediate representation generated by the network. This approach focuses on feature maps at multiple stages of the network, allowing the model to focus on different levels of abstraction within the image.

Both self-attention and attention processes, which are similar strategies, are employed to identify the connections between various elements of an input sequence or collection. While there are some parallels between the two procedures, there are also some significant variances.

A technique called self-attention is used to create a weighted representation of an input sequence or set depending on how similar the items in the sequence or set are to one another. The input sequence or set is divided into the query vector, key vector, and value vector in self-attention. Based on the relationships between the elements in the input sequence or set, these vectors are utilized to create a weight vector that represents the relative importance of each element. The attended representation of the input sequence or set is then represented by computing a weighted sum of the value vectors using the weight vector (figure 1). When determining the score of multiple key and query vectors at the same time, we can replace the key and query vectors with the key and query matrices, K and Q, respectively, in the above equations. Given Q, K, and V, the value of the corresponding query vectors is given by:

$$\text{Attention}(Q, K, V) = V \cdot \text{softmax}(\text{score}(Q, K)) \quad (2)$$

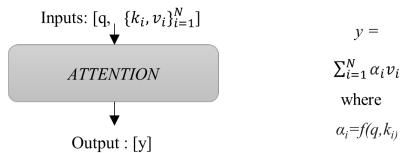


Figure 1: Attention mechanism parameters.

The self-attention processes are similar, except that instead of functioning between the encoder and decoder components, they also operate between the

input and output elements (the present looks at the past and future since the future is yet to be generated) (Jlassi et al., 2021).

Multi-head attention is an attention mechanism module that runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension (Tiwari and Nagpal, 2022).

## 2.2 Transformer

Based on the concept of self-attention, the transformer design enables the model to record relationships between various elements of the input sequence. The input sequence is divided into three vectors called the query vector, key vector, and value vector in the transformer (Duan and Zhao, 2019). These vectors are then used to calculate attention weights for each element in the input sequence. The attended representation of the input sequence is represented by a weighted sum of the values, which is computed using the attention weights (Duan and Zhao, 2019).

Each layer of the transformer architecture is made up of a feedforward neural network and a multi-head self-attention mechanism. The feedforward neural network gives the model non-linearity and flexibility, while the multi-head self-attention mechanism enables the model to attend to different regions of the input sequence at varying degrees of granularity.

The classic recurrent neural networks (RNNs), which were previously employed for many NLP applications, have a number of disadvantages when compared to the transformer design. The transformer, as opposed to RNNs, is able to capture long-range dependencies in the input sequence without the requirement for recurrence or explicit modeling of the input history. It can also be parallelized more easily than RNNs, which speeds up training and makes scaling to bigger datasets simpler.

It is used to parallelize sequential data using an encoder-decoder. It is a deep learning model (i.e., a neural network) of type seq2seq, which has the peculiarity that it uses only the attention mechanism and no RNN or CNN. Based on an encoder-decoder architecture (figure 2).

### 2.2.1 Encoder

The encoder is an important part of the architecture of the transformer encoder-decoder. It is in charge of understanding the input sequence through analysis and representation. A continuous representation of the input, or embedding, is created by the encoder after processing the input sequence. The decoder

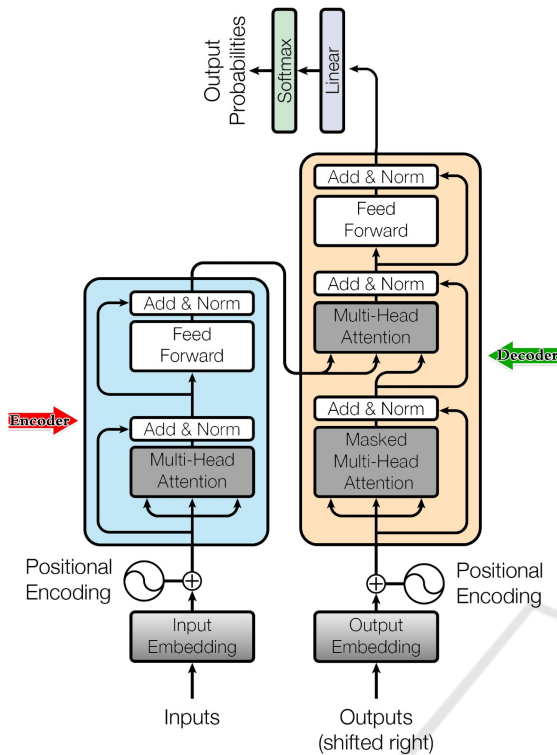


Figure 2: The Transformer – Model Architecture.

then uses these embeddings to produce the output sequence (Han et al., 2022).

A feed-forward neural network and a self-attention mechanism are commonly included in each layer of the transformer encoder architecture. By calculating the dot product of the embeddings, the self-attention mechanism enables the model to assess the relative relevance of various input sequence components. The term "multi-head attention" also applies to this technique (Jdey et al., 2012b).

### 2.2.2 Decoder

The task of the decoder, which contains similar sublayers as the encoder, is to generate text sequences. It has two layers of multi-headed attention: a feed-forward layer with point-wise residual connections and a normalization layer after each sublayer. These sublayers exhibit similar behavior to the encoder's linear layers, but each multi-headed attention layer performs a distinct function (Bhatt et al., 2021). A classifier-like linear layer and a softmax are added as a finishing touch to obtain the word probabilities. The encoder output, which comprises the attention information from the input, and a list of prior outputs are both inputs that the auto-regressive decoder uses. When it produces an end token as an output, the decoder stops decoding (Hcini et al., 2022).

## 3 RELATED WORKS

The ViT architecture has shown promising results in various computer vision tasks, including the classification of satellite images. In particular, the use of ViT in satellite image classification has been studied in a few research papers. In (Bazi et al., 2021), the authors explored the use of ViT for multi-label classification of satellite images. They compared ViT to other state-of-the-art models and found that it outperformed them on several datasets. ViT draws the attention of several researchers on the classification of remote sensing images, such as the authors of (Pei et al., 2019) whose study covers more than 60 recent methods based on transformers for different remote sensing problems in the subfields of remote sensing: Very High Resolution (VHR), hyperspectral (HSI) and Synthetic Aperture Radar (SAR) imagery.

As it is shown in table 1 there are some studies were conducted to evaluate the performance of ViT classifier for image classification.

In 2022 (Mehmood et al., 2022), ViT was applied to two datasets, Sentinel-2 and EuroSAT, achieving an accuracy score of 95% for both datasets.

In 2021 (Li et al., 2022a), a new method called C-Tran based on ViT with RGB input was proposed and evaluated on two datasets, COCO and Visual Genome, with the results showing an accuracy score of 90.1%, a recall score of 65.7%, and an F1 score of 76% for COCO, and an accuracy score of 51.1%, a recall score of 12.5%, and an F1 score of 20.1% for Visual Genome.

In 2020 (Hang et al., 2020), ViT was evaluated on two datasets: Houston 2013 (16 class) with Attention-Aided CNNs, achieving an accuracy score of 90.38% and an F1 score of 90.67%, and Houston 2018 (20 class) with 48 bands achieving an accuracy score of 72.57%. Another dataset, HyRANK (7 class), was evaluated with 176 bands, achieving an accuracy score of 58.55%. Overall, these studies show that ViT can achieve high accuracy scores across different types of datasets and applications, making it a promising tool for future research.

## 4 PROPOSED APPROACH

Our proposed approach is about applying the ViT model without data augmentation using the Keras framework for SAT4 and SAT6 where the ViT model brings several benefits to these two datasets for image classification. They excel in capturing long-range relationships in satellite imagery, offer scalability and transfer learning from large datasets, provide inter-

Table 1: Accuracy obtained for satellite image using ViT Classifier.

Ref	Datasets	Number of images	Classifier	Type of image	Accuracy	Recall	F1
2022 (Mehmood et al., 2022)	<ul style="list-style-type: none"> <li>• Sentinel-2</li> <li>• EuroSAT (10 class)</li> </ul>	<ul style="list-style-type: none"> <li>• 27,000</li> </ul>	<ul style="list-style-type: none"> <li>• ViT</li> </ul>	<ul style="list-style-type: none"> <li>• RGB</li> <li>• 13bands</li> </ul>	95%	–	–
2021(Li et al., 2022a)	<ul style="list-style-type: none"> <li>• COCO</li> <li>• Visual Genome</li> </ul>	<ul style="list-style-type: none"> <li>• 122218</li> <li>• 108077</li> </ul>	C-Tran	RGB	90.1%	65.7%	76%
2021 (Bazi et al., 2021)	<ul style="list-style-type: none"> <li>• Merced (21 class)</li> <li>• AID (30 class)</li> <li>• Optimal31 (31 class)</li> </ul>	<ul style="list-style-type: none"> <li>• 2100</li> <li>• 6600</li> <li>• 1860</li> </ul>	A remote-sensing scene-classification method based on vision transformers	RGB+Nir	<ul style="list-style-type: none"> <li>• 98.49%</li> <li>• 95.86%</li> <li>• 95.56%</li> </ul>	–	–
2020 (Hang et al., 2020)	<ul style="list-style-type: none"> <li>• Houston 2013 (16 class)</li> <li>• Houston 2018 (20 class)</li> <li>• HyRANK (7 class)</li> </ul>	<ul style="list-style-type: none"> <li>• 6700</li> <li>• 6700</li> <li>• 1800</li> </ul>	Attention-Aided CNNs	<ul style="list-style-type: none"> <li>• 144 bands</li> <li>• 48 bands</li> <li>• 176 bands</li> </ul>	<ul style="list-style-type: none"> <li>• 90.38%</li> <li>• 72.57%</li> <li>• 58.55%</li> </ul>	–	<ul style="list-style-type: none"> <li>• 90.67%</li> <li>• 58.02%</li> <li>• 51.61%</li> </ul>

pretable attention maps, and have a track record of top-tier performance. Their adaptability to different patch sizes and availability of pre-trained models make ViT models a compelling choice for improving accuracy and generalization in satellite image classification tasks, provided they are fine-tuned and customized to suit the specific dataset characteristics.

The proposed model starts by defining the input layer for flattened images of size 28x28. Patches are then extracted from the input images using the Patches layer. These patches are encoded into a higher-dimensional representation using the PatchEncoder layer. The model then proceeds with multiple Transformer blocks, each consisting of layer normalization, multi-head self-attention, skip connections, and MLP layers. The final encoded patches undergo layer normalization and are flattened. Dropout is applied to the flattened representation, followed by another MLP layer for feature extraction. The output logits are obtained through a dense layer, and the

model is created using the specified inputs and outputs.

We have conducted a comprehensive examination of the available literature and performed a meticulous investigation to confirm that there are no previous instances where the ViT model has been applied to the Sat4 and Sat6 datasets.

Several key parameters that affect the architecture and behavior of our ViT-based model for image classification are involved. We focus on the patch size hyperparameter in our study, and we try to employ a patch size of 16 applied to an image of size 28\*28.

## 5 EXPERIMENTAL RESULT

### 5.1 Datasets

Table 2 presents data derived from two datasets employed in the assessment of our proposed approach.

Table 3: Exploring the Performance of Vision Transformer Model for Satellite Image Classification.

Datasets	Nb-epochs	Accuracy (%)	Precision (%)	Recall	F1-score	Time(h)
SAT6	• 10	• 96.75	• 95.15	• 93.27	• 93.82	• 00:09:04h
	• 50	• 98.56	• 97.08	• <b>97.44</b>	• 97.23	• 01:13:21h
	• 100	• <b>98.66</b>	• <b>97.80</b>	• 97.27	• <b>97.50</b>	• 01:26:34h
SAT4	• 10	• 96.75	• 92.90	• 92.60	• 92.60	• 01:36:26h
	• 50	• <b>98.61</b>	• 83.37	• 94.75	• 94.09	• 09:04:10h
	• 100	• 94.96	• <b>95</b>	• 94.70	• 94.70	• 09:31:28h

Table 2: Sat4 and Sat6 Characteristics.

Dataset's characteristics	SAT4	SAT6
Total number of images	500000	405000
Training sets	400000	324000
Test sets	100000	81000
Number of class	4 Barren Land, Tree, Grassland, Other Class	6 Building, Barren Land, Tree, Grassland, Road, Water
spatial Resolution	1m	—
Image size	28*28	—
Color depth	8 bits unsigned	—
Bands available	RGB and PIR	—
Approximate location	State of California (united states)	—
size	1.36Go	1.12Go

## 5.2 Results and Discussion

Table 3 shows the results of training the ViT model on Sat4 and Sat6 datasets for different numbers of epochs. The model achieved an accuracy of 98.66% on Sat6 after 100 epochs and 98.61% on Sat4 after 50 epochs of training. The precision, recall, and F1-score for Sat4 were 83.37, 94.75, and 94.09, respectively, while for Sat6, they were 97.80, 97.27, and 97.50, respectively. These results demonstrate that the model was able to correctly classify the image for both datasets, with higher precision and recall achieved on Sat6. The time required to train the model on Sat4 was 9.4 hours, while on Sat6, it was 1.26 hours, indicating that Sat4 took longer to train

due to its larger size and complexity.

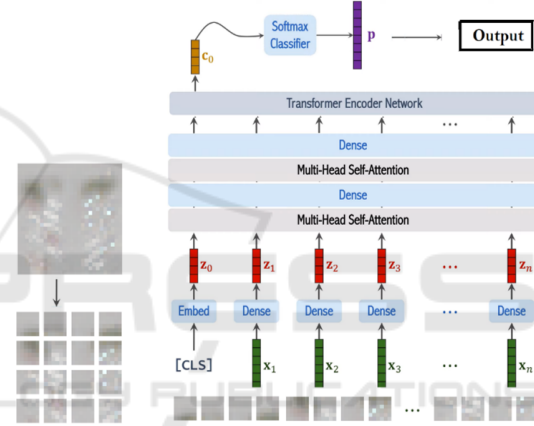


Figure 3: Architecture Vision Transformer.

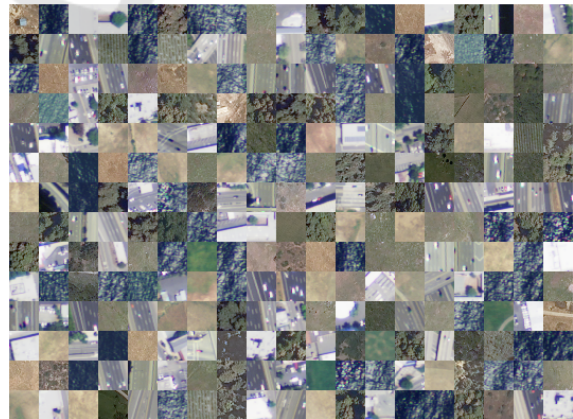


Figure 4: Image from SAT4 and SAT6 datasets.

The hyperparameter range values are shown in table 4. All of the images in our tests and alternatives are sized to 28\*28. The input image's patch size P is

Table 4: The range of hyperparameter values.

Hyperparameter	Description	textbf-Value
Image_size	The height and width of an image in pixels	28 × 28
Patch_size	A small patch means a population of small dimension with greater external influence reaching the inner parts	6
Batch_size	The number of units manufactured in a production run	16
Learning rate	Hyperparameter that controls how much to change the model in response to the estimated error each time the model weights	0.001
Optimizer choice	Stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments with an added method to decay weights (Jdey et al., 2023)	AdamW
Weight decay	Weight decay is a regularization technique that penalizes large weights in the model.	0.0001
Heads	The number of heads refers to the number of parallel self-attention heads in the transformer layers	4
Layer	The number of transformer layers determines the depth of the model	8
Epochs	The number of epochs is an important hyperparameter that determines how long the model trains and how many times it updates its parameters	10,50,100
Dropout rate	The dropout rate is set at 20%, indicating that approximately one out of every five inputs will be randomly omitted during each update cycle	0.2

set to 6, the batch size is set to 16, and the starting learning rate is set to 0.001. With four heads, eight layers, and various values of epochs, the AdamW optimizer is used to train the model deeply. It has a weight decay of 0.0001.

When compared to the works listed in the related works section, our model performs better in terms of accuracy, precision, recall, and training effectiveness. This shows that our model, particularly when applied to the Sat6 dataset, produced state-of-the-art performance in satellite image classification tasks. The particular hyperparameter values we used also point to a well-considered and possibly efficient model setup. The specifics of each dataset and task must be taken into account, though, as performance can vary based on things like the size and complexity of the dataset.

The results suggest that the ViT model was effective in classifying satellite images from both Sat4 and Sat6 datasets, achieving high accuracy and generalization performance. However, the model achieved slightly better results on Sat6, which may be due to its smaller size and simpler features compared to Sat4. The results also highlight the importance of hyperparameter tuning, particularly the number of epochs, in achieving optimal performance of the model.

## 6 CONCLUSION

Our study highlights the effectiveness of the ViT model for satellite image classification, demonstrating its ability to learn meaningful representations of satellite images and generalize well to new data. The results also emphasize the importance of hyperparameter tuning, particularly the number of epochs, in achieving optimal performance of the model. Additionally, our study shows that the ViT model can be applied to different satellite image datasets with varying complexity and features, providing a promising avenue for future research in satellite image classification.

## REFERENCES

- Baig, M. F., Mustafa, M. R. U., Baig, I., Takaijudin, H. B., and Zeshan, M. T. (2022). Assessment of land use land cover changes and future predictions using ca-ann simulation for selangor, malaysia. *Water*, 14(3):402.
- Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., and Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3).
- Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K., and Ghayvat, H. (2021). Cnn

- variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, 10:2470.
- Daud, S. M. S. M., Yusof, M. Y. P. M., Heo, C. C., Khoo, L. S., Singh, M. K. C., Mahmood, M. S., and Nawawi, H. (2022). Applications of drone in disaster management: A scoping review. *Science & Justice*, 62(1):30–42.
- Devi, N. B., Beenarani, B., and Sivanantham, E. (2023). Satellite image detection and classification using hybrid segmentation and feature extraction with enhanced probabilistic neural network. *Earth Science Informatics*, pages 1–12.
- Dimitrovski, I., Kitanovski, I., Kocev, D., and Simidjievski, N. (2023). Current trends in deep learning for earth observation: An open-source benchmark arena for image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:18–35.
- Duan, S. and Zhao, H. (2019). Attention is all you need for chinese word segmentation. *arXiv preprint arXiv:1910.14537*.
- Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110.
- Hang, R., Li, Z., Liu, Q., Ghamisi, P., and Bhattacharyya, S. S. (2020). Hyperspectral image classification with attention-aided cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):2281–2293.
- Hcini, G., Jdey, I., and Ltifi, H. (2022). Improving malaria detection using l1 regularization neural network. *JUCS: Journal of Universal Computer Science*, 285(10).
- Jamil, S., Jalil Piran, M., and Kwon, O.-J. (2023). A comprehensive survey of transformers for computer vision. *Drones*, 7(5):287.
- Jdey, I., Hcini, G., and Ltifi, H. (2023). Deep learning and machine learning for malaria detection: overview, challenges and future directions. *International Journal of Information Technology & Decision Making*.
- Jdey, I., Toumi, A., Dhibi, M., and Khenchaf, A. (2012a). The contribution of fusion techniques in the recognition systems of radar targets.
- Jdey, I., Toumi, A., Khenchaf, A., Dhibi, M., and Bouhlel, M. (2012b). Fuzzy Fusion System for Radar Target Recognition. *International Journal of Computer Applications and Information Technology*, 1(3):136–142.
- Jiang, X., Wang, Y., Liu, W., Li, S., and Liu, J. (2019). Capsnet, cnn, fcn: Comparative performance evaluation for image classification. *Int. J. Mach. Learn. Comput*, 9(6):840–848.
- Jlassi, S., Jdey, I., and Ltifi, H. (2021). Bayesian hyperparameter optimization of deep neural network algorithms based on ant colony optimization. In *Document Analysis and Recognition-ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16*, pages 585–594. Springer.
- Li, G., Chen, X., Li, M., Li, W., Li, S., Guo, G., Wang, H., and Deng, H. (2022a). One-shot multi-object tracking using cnn-based networks with spatial-channel attention mechanism. *Optics and Laser Technology*, 153:108267.
- Li, T., Zhang, Z., Pei, L., and Gan, Y. (2022b). Hashformer: Vision transformer based deep hashing for image retrieval. *IEEE Signal Processing Letters*, 29:827–831.
- Mehmood, M., Shahzad, A., Zafar, B., Shabbir, A., and Ali, N. (2022). Remote sensing image classification: A comprehensive review and applications. *Mathematical Problems in Engineering*, 2022:1–24.
- Pei, Y., Huang, Y., Zou, Q., Zhang, X., and Wang, S. (2019). Effects of image degradation and degradation removal to cnn-based image classification. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1239–1253.
- Scheibenreif, L., Hanna, J., Mommert, M., and Borth, D. (2022). Self-supervised vision transformers for land-cover segmentation and classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1421–1430.
- Slimani, N., Jdey, I., and Kherallah, M. (2023). Performance comparison of machine learning methods based on cnn for satellite imagery classification. In *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 185–189. IEEE.
- Tiwari, D. and Nagpal, B. (2022). Keaht: A knowledge-enriched attention-based hybrid transformer model for social sentiment analysis. *New Generation Computing*, 40(4):1165–1202.
- Xian, T., Li, Z., Zhang, C., and Ma, H. (2022). Dual global enhanced transformer for image captioning. *Neural Networks*, 148:129–141.
- Zu, B., Wang, H., Li, J., He, Z., Li, Y., and Yin, Z. (2023). Weighted residual self-attention graph-based transformer for spectral-spatial hyperspectral image classification. *International Journal of Remote Sensing*, 44(3):852–877.