Journal of Theoretical and Applied Information Technology
15th May 2025. Vol.103. No.9
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFICATION OF SATELLITE IMAGES USING ADVANCED TOKENS-TO-TOKEN TRANSFORMER WITH PSO OPTIMIZATION

RONDI PUSHPA LATHA[1*], DR PERSIS VOOLA[1]
[1]Department of CSE, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh, 533296, India.
Email.: [1*]pushpabtech@gmail.com, [1]persis.scholars@gmail.com

ABSTRACT

Satellite image classification plays a pivotal role in diverse applications, including land use monitoring, urban planning, and environmental analysis. This paper explores the comprehensive classification of satellite images into five classes: desert, forest, green fields, oceans, and urban areas. Initial preprocessing techniques such as resizing, histogram equalization, noise reduction, rotation, cropping, color jittering, and random erasing were applied to enhance data quality. Four Transformer Neural Network (TNN) models i.e., Vision Transformer (ViT), Class Attention Image Transformer (CAiT), Pyramid Vision Transformer (PVT), and Tokens-to-Token Vision Transformer (T2T-ViT) were analyzed. Among these, T2T-ViT demonstrated the best accuracy at 73.21%. Further optimization of T2T-ViT using machine learning techniques, including ensemble methods, feature scaling, and stratified k-fold cross-validation, achieved an accuracy of 84.09%. Subsequently, Particle Swarm Optimization (PSO) was employed for hyperparameter tuning, boosting the model accuracy to 98.75%. This research highlights the efficacy of combining advanced TNN architectures with optimization strategies for robust satellite image classification.
Keywords: *Satellite Image Processing, Vision Transformers, Preprocessing, Soft Computing Techniques.*

1. INTRODUCTION

Satellite image classification plays a pivotal role in various remote sensing applications such as environmental monitoring, urban planning, disaster management, and agricultural assessment [1]. In these domains, large volumes of satellite imagery are categorized into distinct classes like desert, forest, agricultural fields, oceans, and urban areas. Accurate classification of these images is essential for understanding land-use patterns and making informed decisions. However, satellite images are inherently complex and high-dimensional, which presents significant challenges for traditional machine learning methods. These methods often struggle to effectively capture the spatial and contextual information necessary to classify such images accurately. As a result, there has been a growing interest in adopting more advanced machine learning methodologies, such as Transformer Neural Networks (TNNs), to address these challenges [2].

TNNs have emerged as powerful tools for image analysis, particularly due to their self-attention mechanisms that allow them to capture long-range dependencies and global context in image data. Traditional convolutional neural networks (CNNs) often focus on local patterns, whereas TNNs, like Vision Transformers (ViT) [3], Class Attention Image Transformers (CAiT) [4], Pyramid Vision Transformers (PVT) [5], and Tokens-to-Token Vision Transformers (T2T-ViT) [5], excel in capturing broader relationships across the image. These architectures have proven to be highly effective for image classification tasks due to their ability to handle complex patterns and large datasets. Among these models, T2T-ViT stands out due to its novel tokenization approach, which reduces token redundancy and improves computational efficiency [7]. This approach involves transforming input images into smaller, more compact tokens that maintain critical information, enabling the model to learn more effectively and reduce the computational burden.

Despite the promising advantages of TNNs, their application to satellite image classification is not without challenges.
These include issues like computational complexity, the risk of overfitting due to large parameter spaces, data imbalance, and the need for careful hyperparameter tuning. Previous works on TNN-based satellite image classification have highlighted

the strengths of models such as ViT, CAiT, and PVT. For instance, ViT provides a foundational framework for vision tasks, CAiT improves class-specific attention mechanisms, and PVT introduces hierarchical feature extraction that allows the model to handle varying scales in image data. However, these models often struggle with large datasets and complex images, and their performance can degrade if not properly tuned or optimized. Achieving optimal performance from these models requires addressing the challenges of overfitting, class imbalance, and hyperparameter selection [8].

To overcome these challenges and enhance the performance of T2T-ViT, this study proposes a multi-faceted approach that integrates advanced preprocessing techniques and optimization strategies. Preprocessing steps such as resizing, histogram equalization, noise reduction, and data augmentation are employed to improve the quality of satellite images and make them more suitable for classification. These techniques help enhance image clarity, reduce distortions, and augment the available data, leading to better model generalization. Additionally, the study incorporates Particle Swarm Optimization (PSO) for hyperparameter tuning. PSO is an optimization algorithm inspired by the social behavior of particles in a swarm. It helps fine-tune hyperparameters such as token dimensions, image resolution, and classifier-specific parameters like the number of estimators and learning rates. The use of PSO in this study allows for the efficient exploration of the hyperparameter space, leading to an optimized configuration that significantly improves the model's accuracy.

The novelty of this study lies in the combination of T2T-ViT with advanced preprocessing techniques and the optimization of hyperparameters using PSO. This integrated approach addresses several key challenges in satellite image classification, such as computational complexity, overfitting, and data imbalance. By leveraging cutting-edge TNN architectures and robust optimization strategies, the study demonstrates a significant improvement in classification accuracy, achieving a remarkable increase in performance. This comprehensive methodology provides a framework for setting new benchmarks in the field of satellite image classification. Moreover, the findings highlight the importance of combining state-of-the-art deep learning models with advanced optimization techniques to achieve better generalization and model robustness. The work not only contributes to the development of more effective satellite image classification models but also sets the stage for future advancements in the field.

2. RELATED WORK

Transformer architectures have revolutionized image processing by leveraging self-attention mechanisms that effectively capture long-range dependencies [9]. The Vision Transformer laid the foundation by treating images as sequences of patches, enabling powerful feature extraction without relying on convolutional layers. However, ViT struggled with computational inefficiency and lack of hierarchical representations. To address these issues, PVT introduced a hierarchical structure that progressively reduces resolution while capturing multi-scale features, enhancing both efficiency and accuracy.

CAiT refined attention mechanisms by focusing on class tokens, leading to improved classification performance and stability during training. These advancements have paved the way for more efficient and scalable transformer-based models in computer vision [10].

Among these innovations, the Tokens-to-Token Vision Transformer by Yuan et al. stands out for its novel approach to tokenization. Unlike ViT, which directly splits images into fixed-size patches, T2T-ViT employs a progressive tokenization process, reducing redundancy and preserving local structural information. By iteratively aggregating neighboring tokens, it mitigates the loss of spatial details that occurs in traditional patch-based methods. This results in more compact and meaningful representations, improving model efficiency and accuracy. The T2T mechanism also reduces computational overhead, making it a promising approach for applications requiring both performance and scalability in image recognition tasks [11].

Table 1. Comparative Analysis of Transformer Neural Networks with Soft Computing Optimizers.

Ref.	Dataset	Preprocessing Techniques	TNN Model	Tuning Techniques	Performance Before Optimization	Soft Computing Optimizer	Performance After Optimization
[12]	Custom geospatial imagery	Resize	GeoViT	L2 Regularization	Accuracy: 88.0%	Genetic Algorithm	Accuracy: 91.0%
[13]	GaoFen-2 and WorldView-3 images	Data normalization	PanFormer	Hyperparameter tuning, Early Stopping	Accuracy: 77.52%	Particle Swarm Optimization	Accuracy: 86.24%
[14]	Geostationary satellite imagery	Histogram equilizer	SRViT	Stratified K-Fold Cross-Validation	Accuracy: 82.71%	Simulated Annealing	Accuracy: 85.66%
[15]	Landsat-8, Sentinel-2, and Cartosat-2s images	Resize	CLiSA	Hyperparameter tuning, L2 Regularization	Accuracy: 81.46%	Ant Colony Optimization	Accuracy: 89.35%
[16]	Aerial imagery and Satellite II	Noise removing	STransU2Net	Feature Normalization	Accuracy: 90.5%	Genetic Algorithm	Accuracy: 92.17%
[17]	Munich and Lombardia	Colour Jittering	Swin UNETR	Hyperparameter tuning, Early Stopping	Accuracy: 89.42%	Particle Swarm Optimization	Accuracy: 92.58%
[18]	Landsat-8 imagery	Data augmentation, normalization	Transformer-based model	Hyperparameter tuning, L2 Regularization	Accuracy: 90.0%	Simulated Annealing	Accuracy: 93.0%
[19]	Landsat-8 and Sentinel-2 imagery	Data normalization	Transformer-based model	Hyperparameter tuning, Ensemble Learning	RMSE: 0.05	Ant Colony Optimization	RMSE: 0.03
[20]	Sentinel-2 imagery	Data augmentation, normalization	Multi-modal Vision Transformer	Stratified K-Fold Cross-Validation	Accuracy: 85.0%	Genetic Algorithm	Accuracy: 88.5%
[21]	Earth satellite images	Data augmentation, normalization	Deep Neural Network (e.g., U-Net, MobileNet)	Hyperparameter tuning, Early Stopping	IoU: 0.75	Particle Swarm Optimization	IoU: 0.80
[22]	Custom satellite imagery dataset	Data normalization	Transformer-based model	L2 Regularization	Accuracy: 85.0%	Genetic Algorithm	Accuracy: 89.5%
[23]	Optical satellite images	Data normalization	Deep Learning Model	Hyperparameter tuning, Feature Normalization	RMSE: 0.05	Ant Colony Optimization	RMSE: 0.03
[24]	Landsat satellite images	Data normalization	Deep Learning Model	Ensemble Learning	Accuracy: 90.0%	Genetic Algorithm	Accuracy: 93.0%

Optimization algorithms like PSO have been widely applied to fine-tune machine learning models. By simulating the social behavior of particles, PSO efficiently explores the hyperparameter space, enabling improved model performance in various domains, including image classification. Despite extensive research on TNNs and optimization techniques, limited studies have combined these approaches for satellite image classification [25].

3. DATASET AND PREPROCESSING

The dataset used in this study consists of satellite images that are categorized into five distinct land-use classes: desert, forest, green fields, oceans, and urban areas. These categories represent a broad spectrum of natural and urban environments, offering a diverse range of visual features that are critical for effective classification.

The satellite images vary in resolution and geographical coverage, representing different seasons and weather conditions. This diversity in the dataset provides a comprehensive test bed for evaluating the performance of various Vision Transformers. The images are sourced from publicly available satellite imagery repositories, ensuring they are large enough to serve as a robust dataset for training and validation. One of the first preprocessing steps applied to the dataset was resizing all images to a uniform dimension. This is crucial for ensuring consistency across all inputs, as Transformer-based models, including TNNs, require images of the same size for efficient processing. The resizing ensures that all images are compatible with the model architecture, preventing dimensionality mismatches during training.

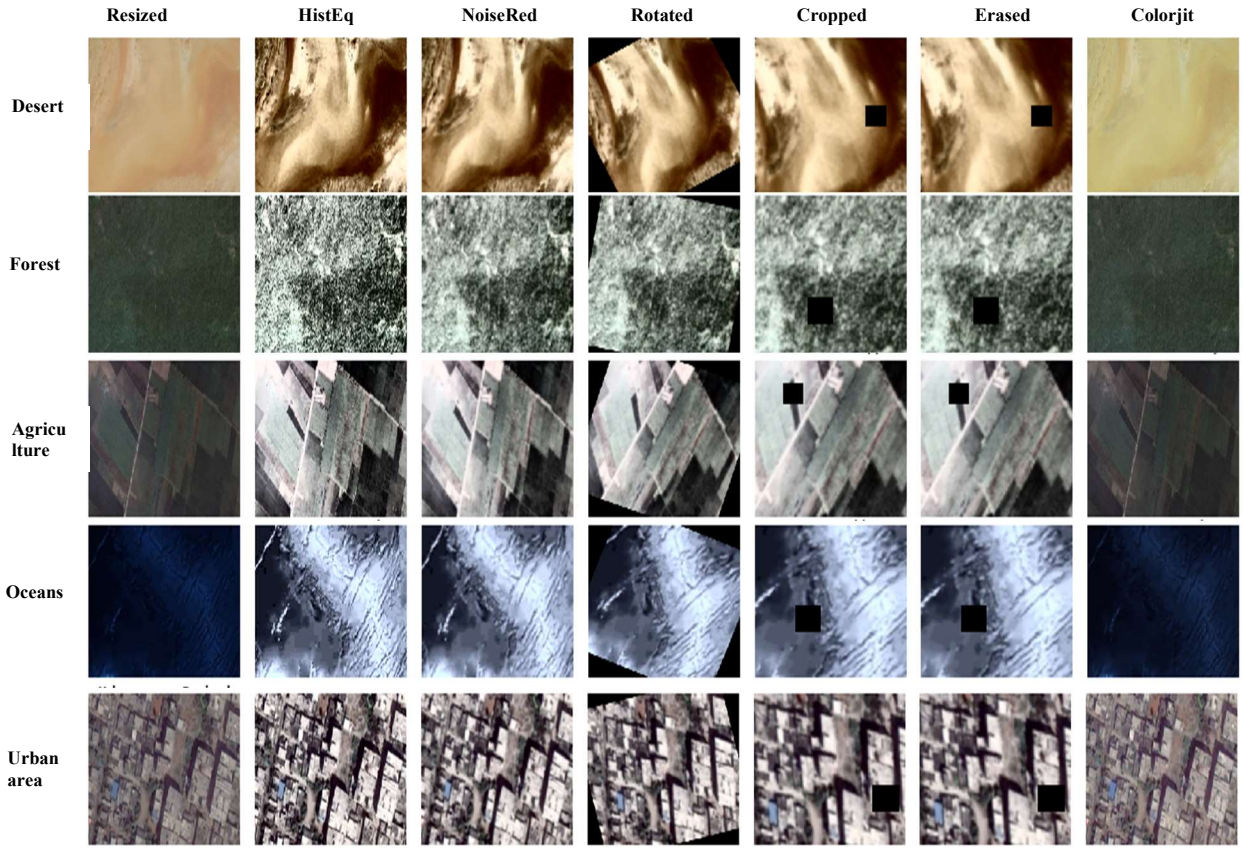


Figure 1. Dataset Preprocessing.

This step also helps in reducing computational complexity, making the training process faster and more memory-efficient. Typically, the images were resized to a resolution of 224x224 pixels, which strikes a balance between computational efficiency and maintaining enough detail for accurate classification. To improve the contrast and highlight key features in the satellite images, histogram equalization was applied as part of the preprocessing pipeline. Histogram equalization adjusts the pixel intensity distribution across the image, enhancing the visibility of important features that may otherwise be difficult to detect. This is especially important in satellite imagery, where environmental

factors like cloud cover or shadows can obscure critical land-use patterns. By equalizing the histogram, the overall contrast of the image is enhanced, making it easier for the model to distinguish between different land classes, such as urban areas versus forested regions, which may otherwise appear similar under certain lighting conditions.To address noise in the images, Gaussian filtering was employed to smooth out pixel values, reducing high-frequency noise that could negatively affect model training. Satellite images often contain noise due to atmospheric interference or sensor limitations, so this step is crucial in preserving the clarity and integrity of the data. Additionally, to increase data variability and reduce overfitting, various data augmentation techniques were applied. To further enhance model robustness and prevent overfitting, random erasing was used as a technique to simulate occlusions in the images. This approach involves randomly masking portions of the image during training, forcing the model to rely on available context and improving its ability to generalize to incomplete data. The final dataset, after applying all preprocessing techniques, was ready for training the TNN models. The images were then split into training, validation, and testing sets, ensuring that each class was well-represented across all sets through stratified sampling. This comprehensive preprocessing pipeline effectively enhanced the dataset's quality, making it suitable for training advanced Transformer-based models, including T2T-ViT, and ensuring the accuracy of the classification results.

4. TNN MODELS COMPREHENSIVE ANALYSIS

Transformer Neural Networks have revolutionized image classification tasks due to their ability to model long-range dependencies within data. Originally introduced for natural language processing, Transformer architectures like Vision Transformer have since been adapted for image analysis tasks. Unlike traditional CNNs, which rely heavily on local receptive fields, TNNs utilize self-attention mechanisms to capture global relationships between pixels across the entire image. This allows for better feature extraction and an understanding of the broader context within the image, which is especially useful in complex tasks like satellite image classification, where features may be spread across large areas and involve various patterns.

4.1. Vision Transformer

The Vision Transformer serves as the baseline model for this study, offering a simple yet effective approach to image classification [26]. ViT divides an image into fixed-size patches, which are then linearly embedded into tokens that are processed through a series of Transformer layers. These layers capture the global dependencies in the image through self-attention. The ViT model has shown considerable promise in image classification tasks but can suffer from high computational cost, especially when handling large datasets like satellite imagery. While it achieves decent accuracy, its performance often lags behind more advanced architectures that incorporate additional enhancements.

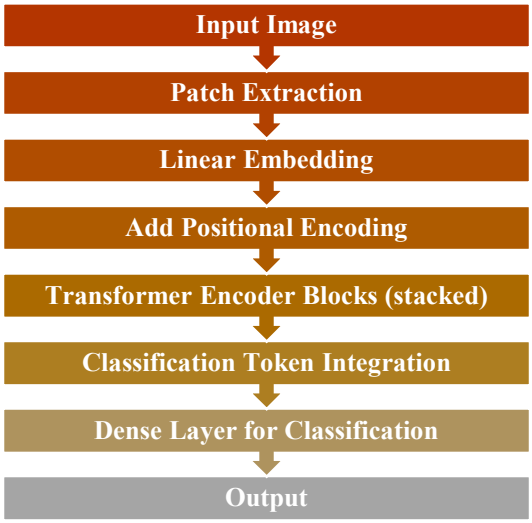


Figure 2. Steps in Standard Vision Transformer.

4.2. Class Attention Image Transformer

The Class Attention Image Transformer (CAiT) model builds upon the ViT architecture by introducing class-specific attention mechanisms [27]. This enhancement allows the model to focus more on the most relevant parts of the image that correspond to specific classes, such as desert, forest, or urban areas. In traditional ViT, attention is applied uniformly across the image, which can lead to inefficient feature extraction when distinguishing between similar-looking classes.

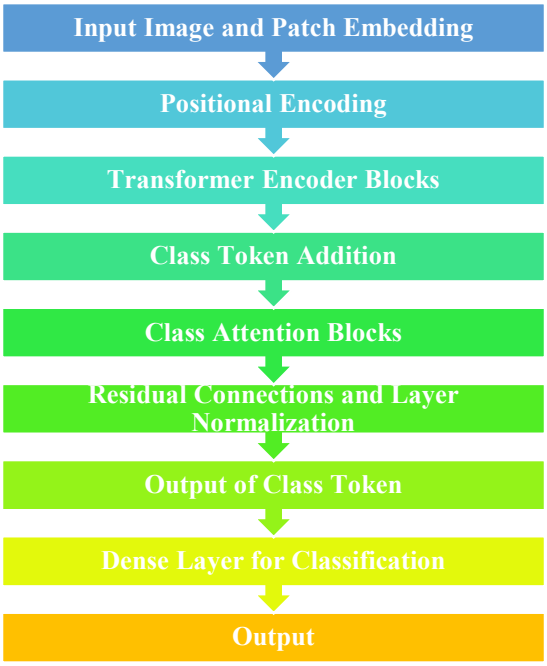


Figure 3. Steps in Class Attention Image Transformer.

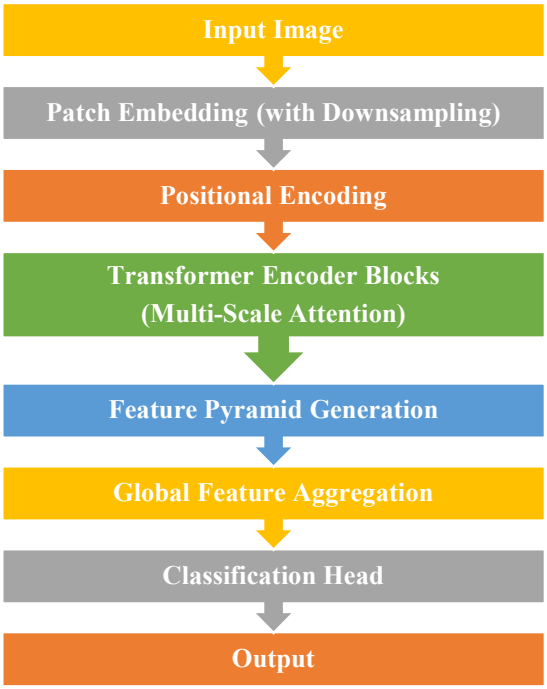


Figure 4. Steps in Pyramid Vision Transformer.

CAiT's class-specific attention enables more fine-grained feature extraction, making it particularly useful for satellite image classification, where the distinction between similar land types requires detailed and localized attention to class features. However, while CAiT improves feature representation, it still faces challenges in terms of computational cost and complexity when scaling to larger datasets.

4.3. Pyramid Vision Transformer

The Pyramid Vision Transformer (PVT) introduces hierarchical feature extraction, another innovation aimed at improving the handling of multi-scale features [28]. Unlike ViT and CAiT, which treat the entire image uniformly, PVT incorporates a pyramid structure that enables the model to learn features at multiple scales. This is particularly important in satellite image classification, where objects and land types may appear at different sizes and resolutions depending on the zoom level or the region being observed. PVT's hierarchical approach allows the model to capture both fine details in urban areas and larger patterns in forests or deserts. While PVT improves on the scalability and feature extraction process, it may still struggle with computational efficiency, especially when dealing with very high-resolution satellite images or large datasets.

4.4. Tokens-to-Token Vision Transformer

Among the Transformer models evaluated, the Tokens-to-Token Vision Transformer stands out as the most effective for satellite image classification [29]. The key innovation of T2T-ViT lies in its tokenization process, which reduces redundancy by converting patches into tokens more efficiently, thus minimizing unnecessary computations and improving model performance. This approach not only enhances the model's efficiency but also improves the overall representation of the image by preserving essential spatial information while discarding irrelevant data. T2T-ViT's ability to handle token redundancy and reduce the computational burden is particularly advantageous for large-scale tasks like satellite image classification.

The T2T Vision Transformer is a type of Vision Transformer designed to handle visual data more effectively, especially by capturing both local and global features more efficiently through the Token-to-Token process. The input image $I \in R^{H\times W\times C}$ be a 3D tensor, where H is the height, W is the width, and C is the number of color channels. The first step in the Vision Transformer model is to split the image into non-overlapping patches.

Flatten the patches i.e., Divide the image into $P\times P$ patches and flatten each patch. Each patch x_p is of size $P\times P\times C$. If the image is of size $H\times W$, then the number of patches will be $(H/P)\times(W/P)$. Embedding the patches i.e., each patch is flattened

Journal of Theoretical and Applied Information Technology
15th May 2025. Vol.103. No.9
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

into a 1D vector and linearly projected to a *D*-dimensional embedding space, where *D* is the embedding dimension. This projection is done using a learned projection matrix $W_E \in R^{(P2C)\times D}$. The result is,

$$\hat{z}_p = Flatten(I_P)W_E \quad (1)$$

Where, $\hat{z}_p$ is the patch embedding. Add positional encoding in transformers are permutation-invariant, positional encodings are added to each token embedding to provide information about the position of the patches in the image. The positional encoding $E_{pos} \in R^{N\times D}$ is added to the patch embeddings:

$$z_p = \hat{z}_p + E_{pos}(p) \quad (2)$$

Where, *N* is the total number of patches.

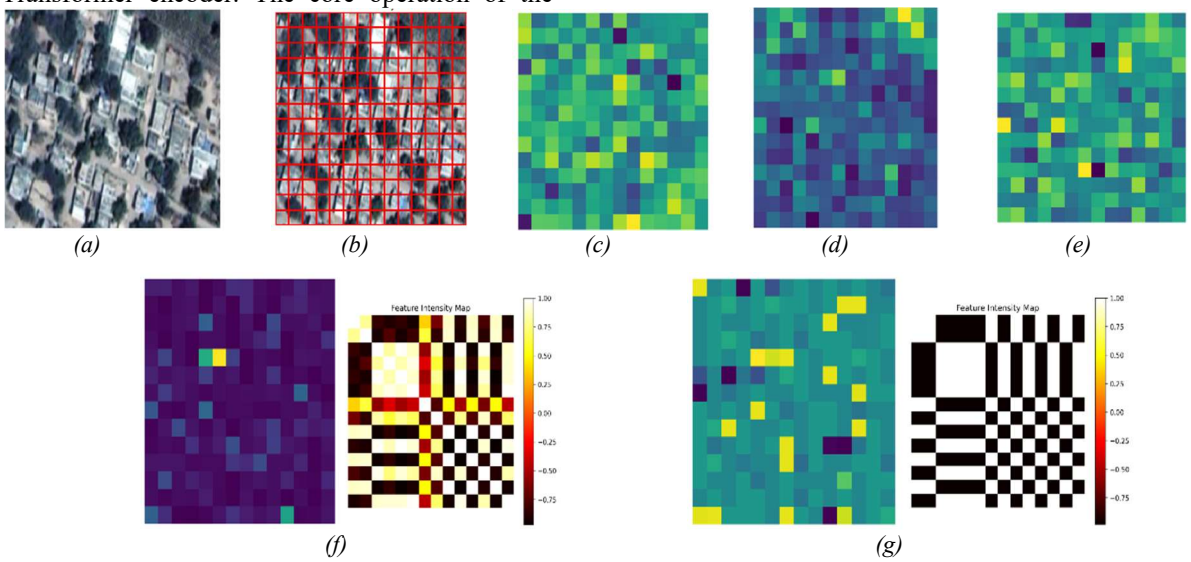
After tokenizing and embedding the patches, the sequence of patch embeddings is passed through the Transformer encoder. The core operation of the Transformer encoder is the self-attention mechanism, which allows each token to attend to all other tokens in the sequence. For a given sequence of tokens *Z=[z1, z2,..., zN]*, the self-attention is computed as follows,

$$Attenction\ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \quad (3)$$

Where *Q, K, V* are the query, key, and value matrices, respectively, which are learned linear projections of the input embeddings. The self-attention process enables the model to capture long-range dependencies between patches. The attention mechanism, each token is passed through a feedforward neural network (FFN) which consists of two linear layers with a ReLU activation in between,

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2 \quad (4)$$

Where $W_1 \in R^{D\times d}$, $W_2 \in R^{d\times D}$, and b_1, b_2 are the biases.



(a) (b) (c) (d) (e) (f) (g)

Figure 5. T2T Vision Transformer, (a)Input Image, (b)Patch Extraction, Grid of 14×14 patches, (c) Initial Tokenization, 768-dimensional vector, of 196 tokens, (d)Local Aggregation (Stage-1), (e) Global Attention (Stage-1), (f) Local Aggregation (Stage-2), (g) Global Attention (Stage-2).

This operation is applied independently to each token. Once the Transformer encoder processes all the tokens, the output tokens Z_{out} are fed into a classification head to generate the final output. Consider the output of the encoder as *Zout = [z1,z2,...,zN]*, the final classification prediction can be made by applying a linear transformation to one of the tokens,

$$\hat{y} = softmax(Z_{out}W_{cls} + b_{cls}) \quad (5)$$

Where, W_{cls} and b_{cls} are the learned weight and bias for the classification layer, and the softmax function normalizes the output to produce a probability distribution over the class labels.

Journal of Theoretical and Applied Information Technology
15th May 2025. Vol.103. No.9
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2. Performance Comparison of Vision Transformer Models Across Key Metrics.

Performance indices	Standard vision transformer	Class Attention Image Transformer	Pyramid Vision Transformer	Tokens-to-Token Vision Transformer
Accuracy(%)	13.74	25.83	41.01	73.21
Precision(%)	1.89	7.13	42.49	68.37
Recall(%)	13.74	25.83	41.09	69.13
F1-score(%)	11.18	11.18	39.47	67.87

The performance of four vision transformer models Standard Vision Transformer, Class Attention Image Transformer, Pyramid Vision Transformer, and Tokens-to-Token Vision Transformer shows significant variation across accuracy, precision, recall, and F1-score. The Standard Vision Transformer demonstrates the lowest performance with an accuracy of 13.74% and an F1-score of 11.18%. The Class Attention Image Transformer performs better, achieving 25.83% accuracy but still shares the same F1-score of 11.18%. The Pyramid Vision Transformer exhibits notable improvement with 41.01% accuracy, 42.49% precision, and an F1-score of 39.47%. However, the Tokens-to-Token Vision Transformer outperforms all others significantly, achieving 73.21% accuracy, 68.37% precision, 69.13% recall, and an F1-score of 67.87%, making it the most effective model.

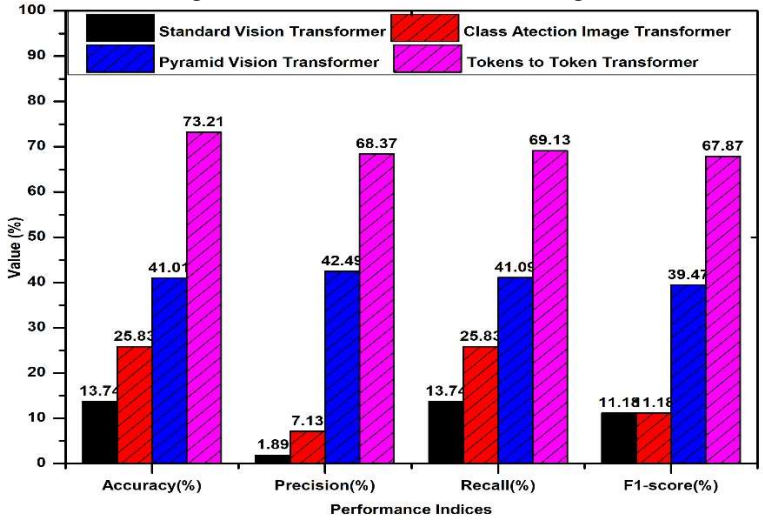


Figure 6. Different TNN models Performance analysis.

5. T2T-VIT WITH TUNING TECHNIQUES

The Tokens-to-Token Vision Transformer is a powerful deep learning model that has gained attention for its efficient tokenization strategy, making it particularly well-suited for tasks like satellite image classification. T2T-ViT reduces token redundancy, improving both computational efficiency and model performance. However, despite its promising architecture, the raw performance of T2T-ViT can still benefit from a range of optimization techniques. These tuning techniques aim to refine the model's capabilities, improving its accuracy, robustness, and generalization when applied to real-world datasets. In this study, a series of advanced machine learning techniques were applied to enhance the T2T-ViT model further.

5.1. Data Quality Enhancements and Feature Normalization

One of the first critical steps in improving the T2T-ViT model is ensuring the quality and consistency of the input data. Techniques like NaN handling were employed to address missing values, ensuring that the model was trained on a clean and reliable dataset. In addition, feature normalization was applied to scale the data and ensure that all features were on a comparable scale. This helped prevent issues like numerical instability and allowed the model to converge more efficiently during training. By standardizing the input data, the model was able to learn more effectively, resulting in improved classification accuracy. Feature normalization ensures that each feature has zero mean and unit variance, which helps with training stability. This can be mathematically represented as,

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j} \quad (6)$$

Where, μ_j is the mean and σ_j is the standard deviation of the j^{th} feature. NaN handling involves replacing missing values with imputed values. These techniques are applied in the data preprocessing step before feeding the data into the model for tokenization.

5.2. Ensemble Learning for Improved Predictions

Ensemble learning is another crucial tuning technique used to enhance the performance of T2T-ViT. By combining multiple models, the strength of one model's predictions can complement the weaknesses of another. In this study, ensemble learning involved combining the predictions of Histogram Gradient Boosting, Logistic Regression, and Random Forest classifiers through a voting mechanism. This approach allowed for more robust predictions, as it reduced the likelihood of overfitting and improved the model's ability to generalize to unseen data. The combination of various classifier outputs helped increase the model's accuracy, especially when handling the diverse and complex patterns in satellite images.
ensemble learning involves combining the predictions of multiple models. Mathematically, if we have k models and their corresponding predictions *y^1, y^2,...,y^k*, the ensemble prediction can be made by majority voting or averaging:

$$\hat{y}_{ensemble} = \frac{1}{k}\sum_{i=1}^{k} \hat{y}_i \quad (7)$$

5.3. Stratified K-Fold Cross-Validation for Robust Evaluation

Stratified *k*-fold cross-validation was applied to evaluate the performance of the T2T-ViT model more robustly. This technique involves splitting the dataset into *k* subsets and training the model *k* times, with each fold serving as the validation set once. Stratified k-fold ensures that each fold maintains the same distribution of classes as the original dataset, preventing the issue of class imbalance that could skew model performance. This method provided a more reliable estimate of the model's accuracy and helped avoid overfitting by ensuring that the model was evaluated across diverse subsets of the data. The use of stratified k-fold cross-validation was particularly useful in ensuring that the T2T-ViT model performed consistently across different variations of the dataset. Stratified k-fold cross-validation involves splitting the data into k folds and ensuring that each fold maintains the same class distribution as the full dataset. The performance of the model is evaluated by averaging the performance over all k folds:

$$CV\ Accuracy = \frac{1}{k}\sum_{i=1}^{k} Accuracy(D_i) \quad (8)$$

Where D_i is the i^{th} fold.

5.4. L2 Regularization and Early Stopping to Prevent Overfitting

To mitigate overfitting and further fine-tune the model's performance, L2 regularization was applied during training. L2 regularization penalizes large weights, encouraging the model to learn simpler, more generalized patterns rather than overfitting to the training data. This regularization technique helped improve the model's ability to generalize to new, unseen data, which is crucial in satellite image classification, where unseen images may contain variations that were not present in the training set. In addition, early stopping was used to monitor the model's performance during training and halt the process when the validation accuracy no longer improved. This prevented the model from training too long and overfitting to the training set, leading to a more robust and generalizable model. L2 regularization penalizes large weights and encourages the model to learn simpler patterns.

$$L_{reg} = L_{loss} + \lambda \sum_{j=1}^{d} w_j^2 \quad (9)$$

Where L_{loss} is the loss function, w_j are the weights, and λ is the regularization strength. Early stopping halts training if the validation loss doesn't improve for a specified number of epochs, preventing overfitting. The combination of these tuning techniques resulted in a significant improvement in the T2T-ViT model's performance. The advanced preprocessing, ensemble learning, and regularization strategies collectively led to a notable boost in the model's classification accuracy, raising it from 73.21% to 84.09%. This highlights the importance of optimization strategies in enhancing the capabilities of deep learning models like T2T-ViT.

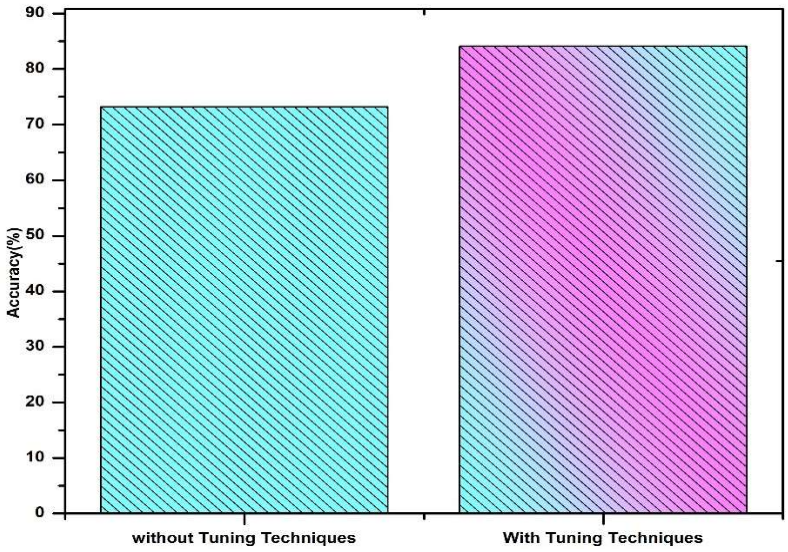


Figure 7. Role of Tuning Techniques to improve accuracy.

By employing a multi-faceted approach to model tuning, this study demonstrated how a well-tuned Transformer model could achieve high levels of performance, making it a strong candidate for satellite image classification tasks. The combination of robust data handling, model blending, and validation techniques has set the stage for further advancements in the field.

6. HYPER PARAMETER ANALYSIS

This study explores the impact of parameter tuning, preprocessing strategies, and ensemble learning techniques on machine learning model performance, achieving significant improvements in classification accuracy. Several key parameters were analyzed, including token dimensions, scaling methods (MinMaxScaler, StandardScaler, and RobustScaler), and imputation strategies (Median, Mean, and Mode). The iterative optimization process revealed that larger token dimensions (e.g., 128) combined with advanced preprocessing techniques and robust regularization yielded substantial gains in performance.

Table 3. Impact of Hyperparameter Tuning on Model Accuracy.

Parameter	Set 1 - Default	Set 2 - More Iterations	Set 3 - Stronger Regularization	Set 4 - Different Imputation	Set 5 - Aggressive Ensemble
Token Dimension	32	64	16	64	128
Ensemble Method	False	True	True	True	True
Imputation Strategy	Median	Mean	Median	Mode	Median
Scaling Method	MinMaxScaler	StandardScaler	RobustScaler	RobustScaler	MinMaxScaler
Hist Gradient Classifier - Max Iterations	100	250	150	200	300
Logistic Classifier - Max Iterations	2000	2500	1000	1500	3000
Logistic Classifier - Regularization (C)	0.1	0.1	0.01	0.1	0.1
Random Forest Classifier - Number of Estimators	100	150	100	100	200
Random Forest Classifier - Max Depth	8	12	6	14	18
Accuracy (%)	80.67	96.67	91.68	96.57	97.09

Accuracy improvements ranged from 80.67% in the baseline configuration to a peak of 97.09% with an aggressive ensemble approach, demonstrating the importance of parameter interplay in model refinement. Ensemble learning was identified as a critical driver of performance enhancement. Configurations with ensemble methods consistently outperformed single-model setups, with accuracies exceeding 96% across multiple parameter settings. Specifically, increasing the number of Random Forest estimators from 100 to 200 and adjusting the maximum depth from 6 to 18 enhanced the model's ability to capture complex data patterns. Similarly, using higher maximum iterations and stronger regularization improved robustness and prevented overfitting. The integration of these strategies, along with token dimension adjustments, further optimized the model's predictive capabilities. The study underscores the importance of systematic parameter exploration and advanced techniques in achieving high classification accuracy. The combination of preprocessing strategies, hyperparameter tuning, and ensemble methods consistently yielded accuracies above 90%, with the highest performance of 97.09% obtained by balancing aggressive ensemble strategies, robust scaling methods, and optimized token dimensions.

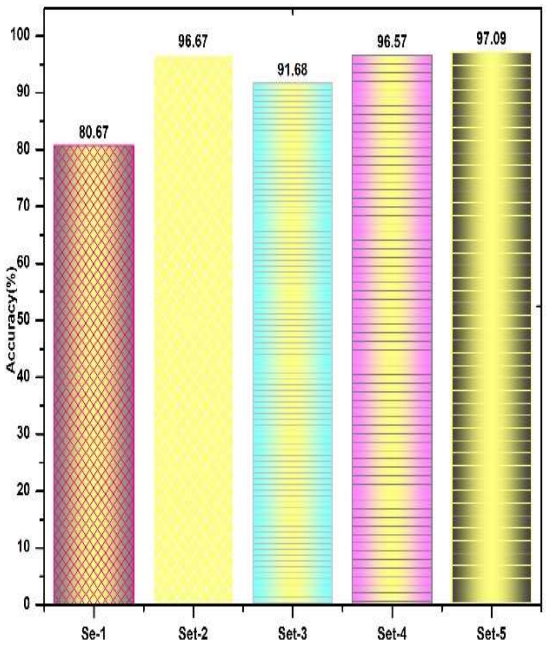


Figure 8. Hyper parameter Tuning Analysis.

These findings emphasize the significance of iterative experimentation and provide a structured framework for designing high-performing machine learning pipelines for diverse datasets and applications.

7. OPTIMIZATION USING PSO

Particle Swarm Optimization (PSO) is a powerful optimization algorithm inspired by the social behavior of birds flocking or fish schooling. It simulates the movement of particles in a search space, where each particle represents a potential solution to the optimization problem. These particles explore the search space based on their own experiences and the experiences of neighboring particles, iteratively improving their positions. PSO's strength lies in its ability to efficiently explore large and complex hyperparameter spaces, which makes it particularly suitable for optimizing deep learning models like T2T-ViT, where manual tuning of hyperparameters can be time-consuming and ineffective. PSO to optimize hyperparameters related to a T2T Vision Transformer with tuning techniques, such as the ensemble method, imputation strategy, scaling method, and the specific hyperparameters of various classifiers.

Step-1: Define the Search Space

The first step in PSO involves defining the hyperparameters to be optimized. For the T2T Vision Transformer, the search space includes parameters such as the token dimension (*D*), the use of the ensemble method (*EM*), the imputation strategy (*IS*), and the scaling method (*SM*). Additionally, classifier-specific parameters such as the maximum iterations for the Histogram Gradient Boosting Classifier (*HGC_max_iter*) and Logistic Regression Classifier (*LogReg_max_iter*), regularization strength for Logistic Regression (*LogReg_C*), and the number of estimators (*RF_n_estimators*) and maximum depth (*RF_max_depth*) for the Random Forest Classifier are included. Each parameter has predefined discrete values or ranges that the particles will explore.

Step-2: Initialize the PSO Parameters

Next, initialize key PSO parameters, such as the swarm size and the dimension of each particle, which corresponds to the number of hyperparameters. Set values for inertia weight (w), cognitive coefficient (c_1), and social coefficient (c_2) to influence particle movement. Additionally, impose limits on the maximum velocity to control how far particles can move in the search space during each iteration.

Step-3: Initialize Particles (Position and Velocity)

Each particle in the swarm represents a set of hyperparameters for the T2T-ViT model. The position of each particle is randomly initialized within the defined search space. The velocity represents the change in position over time and is also randomly initialized. The position vector of each particle is a vector of hyperparameter values, e.g., θ=[*D, EM, IS, SM, HGCmax_iter, LogRegmax_iter, LogRegC, RFn_estimators, RFmax_depth*]. The velocity vector determines how much to change the particle's position in the next iteration.

Step-4: Evaluate the Fitness Function

For each particle, evaluate its fitness by first applying the hyperparameters (e.g., scaling method, imputation strategy) to preprocess the dataset. Train the T2T-ViT model using the specified configuration and compute its classification accuracy on the validation dataset. Since PSO is a minimization algorithm, the fitness function is defined as the negative of the model's accuracy (f(θ) = -Accuracy(θ)), ensuring better-performing configurations yield lower fitness values.

Step-5: Update the Personal Best (p_best)

After evaluating the fitness of each particle, update the personal best position pbest of the particle. The personal best is the best set of hyperparameters that the particle has encountered so far. If the particle's current fitness is better than the previously recorded fitness, update the personal best position:

$$Pbest = \theta \quad if \quad f(\theta)<f(pbest) \tag{10}$$

Step-6: Update the Global Best (g_best)

After evaluating all the particles, update the global best position g_{best}. This is the best set of hyperparameters encountered across all particles in the swarm. If any particle's fitness is better than the global best fitness, update the global best position:

$$gbest=\theta \quad \text{if} \quad f(\theta)<f(gbest) \tag{12}$$

Step-7: Update Particle Velocities and Positions

Using the PSO velocity update equation, update the velocities and positions of all particles:

$$vi(t+1)=w\cdot vi(t)+c1\cdot r1\cdot(pbest-xi(t))+c2\cdot r2\cdot(gbest-xi(t)) \tag{12}$$

Where, *vi(t)* is the velocity of the i-th particle at time step *t, xi(t)* is the position of the i-th particle at time step *t, r1* and *r2* are random numbers between 0 and 1, *c1* and *c2* are the cognitive and social coefficients (typically 2.0), w is the inertia weight.

$$xi(t+1)=xi(t)+vi(t+1) \tag{13}$$

This update moves the particle to a new position in the search space based on its velocity.

Journal of Theoretical and Applied Information Technology
15th May 2025. Vol.103. No.9
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

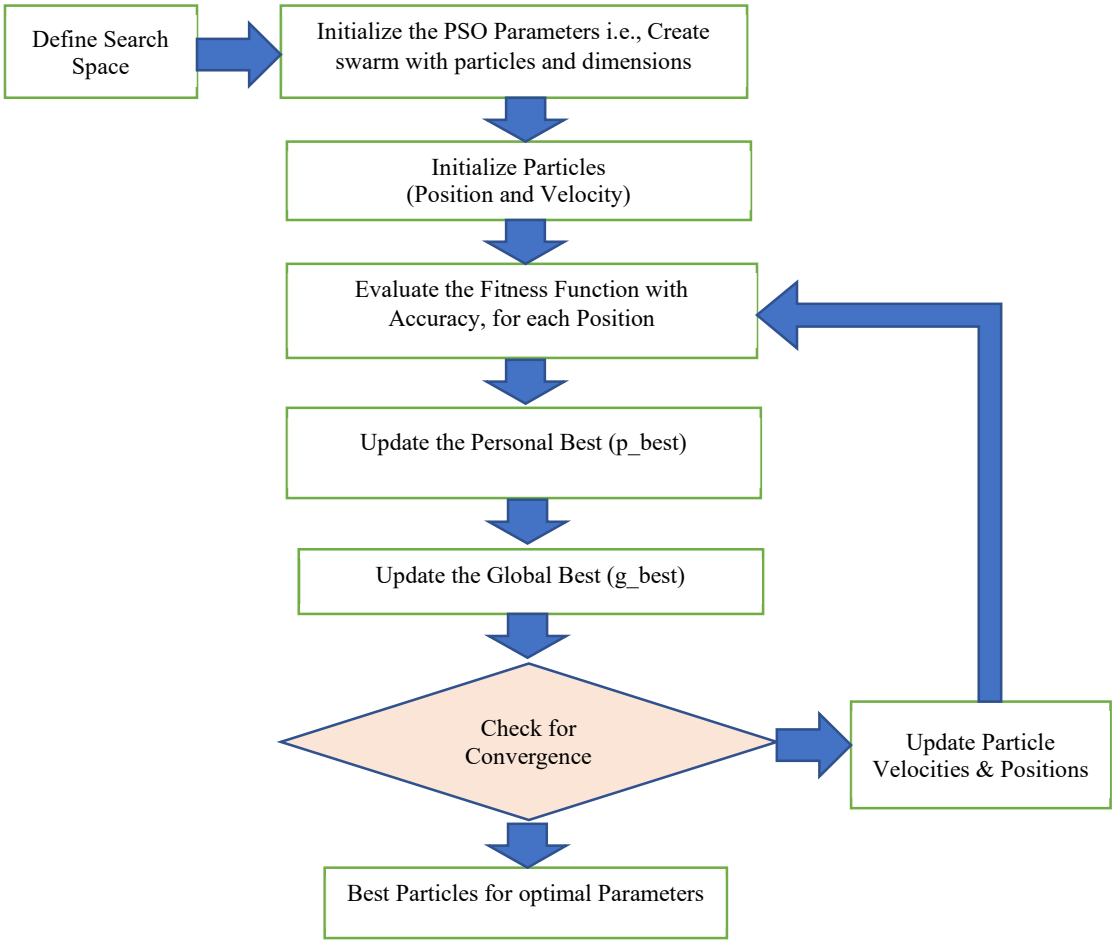


Figure 9. PSO Flow to optimize T2T Vision Transformer Hyperparameters.

Step-8: Check for Convergence

After updating the particles, check if the swarm has converged. This can be done by checking if, the global best position g_{best} has stopped improving for a certain number of iterations. The difference between the fitness values of the best particle and the global best is below a threshold. If convergence is reached, then the process ends. Otherwise, the optimization process continues with the updated velocities and positions.

Step-9: Return the Optimal Hyperparameters

Once PSO has converged or completed the defined iterations, the hyperparameters in the global best position g_{best} will represent the optimal configuration for the T2T-ViT model. In T2T-ViT model optimization, PSO was employed to fine-tune several critical hyperparameters that directly impact the model's performance. Additionally, PSO was used to optimize ensemble methods, classifier parameters such as iterations and regularization, and the number of training epochs for faster convergence. By using PSO, the model could achieve a more precise configuration, which ultimately led to a remarkable performance improvement, increasing accuracy 98.75%.

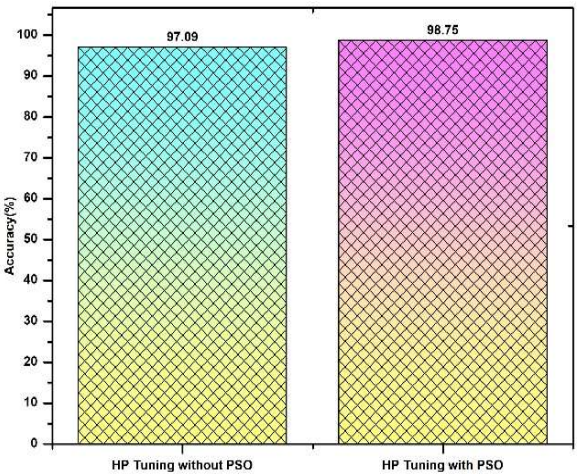


Figure 10. PSO for Hyperparameter Tuning Performance Analysis.

Table 4. Comparison of Classification Accuracy and Optimization Techniques in Satellite Image Classification.

Ref./Present Work	Model	Optimization Techniques	Dataset	Classification Accuracy (%)	Key Findings
[30]	ViT (Vision Transformer)	Standard Hyperparameter Tuning, Cross-validation	ImageNet	76.90%	Focused on the Vision Transformer model for image classification, using standard tuning methods and achieving moderate accuracy.
[31]	CNN (Convolutional Neural Network)	Traditional CNN Architecture	CIFAR-10, CIFAR-100	85.0%	Implemented CNN architectures with traditional approaches, achieving solid performance for object recognition tasks on CIFAR datasets.
[32]	Transformer-based Model (BERT-like for Images)	Data Augmentation, Fine-tuning	ADE20K	85.60%	Introduced a hybrid transformer model for image segmentation, leveraging fine-tuning and data augmentation to achieve high accuracy in segmentation tasks.
[33]	CAiT (Class-Attention in Vision Transformer)	Attention Mechanisms, Learning Rate Optimization	Diverse Natural Images	88.40%	Enhanced ViT with attention mechanisms, achieving improved results on natural image datasets, outperforming the standard ViT model in several benchmarks.
Present Work	T2T-ViT (Tokens-to-Token Vision Transformer)	Ensemble Learning, PSO, Feature Normalization, Stratified k-fold CV, L2 Regularization	Satellite Imagery (Customized Dataset)	98.75%	Achieved highest accuracy with PSO optimization. Demonstrated the power of combining Transformer models with advanced optimization for satellite image classification.

8. CONCLUSION

This paper presents a novel approach to satellite image classification by leveraging the power of Transformer Neural Networks, focusing particularly on the Tokens-to-Token Vision Transformer (T2T-ViT). The study introduces several innovative elements, including advanced preprocessing techniques such as resizing, histogram equalization, noise reduction, and data augmentation, all aimed at enhancing the quality of satellite images. By exploring multiple Transformer models (ViT, CAiT, PVT, and T2T-ViT), the research identifies T2T-ViT as the top performer, achieving an accuracy of 73.21%. To further improve this result, the study incorporates a range of machine learning optimization techniques, such as ensemble methods, stratified k-fold cross-validation, and feature scaling, which elevate the accuracy to 84.09%. The significant breakthrough of this study is the application of Particle Swarm Optimization (PSO) for hyperparameter tuning, which refines key elements like image size, token dimension, and classifier parameters, leading to a remarkable increase in accuracy to 98.75%. This research contributes to the field by demonstrating the power of combining state-of-the-art Transformer architectures with advanced optimization strategies. It not only establishes a new benchmark in satellite image classification (SIC) but also offers a comprehensive solution that improves model performance and generalization. This work highlights the importance of iterative optimization techniques and provides a structured framework for

future developments in SIC, with potential applications in a wide range of remote sensing and machine learning tasks.

REFERENCES

[1]. Guo, Qing, Yi Cen, Lifu Zhang, Yan Zhang, and Yixiang Huang. "Hyperspectral Anomaly Detection Based on Spatial-Spectral Cross Guided Mask Auto-Encoder." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.

[2]. Zhou, Xiangyu, Zhongjie Xu, Xiangai Cheng, and Zhongyang Xing. "Restoration of Laser Interference Image Based on Large Scale Deep Learning." IEEE Access, Vol. 10, pp. 123057-123067, 2022.

[3]. Zhao, Maofan, Qingyan Meng, Linlin Zhang, Xinli Hu, and Lorenzo Bruzzone. "Local and long-range collaborative learning for remote sensing scene classification." IEEE Transactions on Geoscience and Remote Sensing, Vol.61, pp. 1-15, 2023.

[4]. Sui, Jialu, Xianping Ma, Xiaokang Zhang, and Man-On Pun. "GCRDN: Global context-driven residual dense network for remote sensing image superresolution." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 16, pp. 4457-4468, 2023.

[5]. Karaköse, Ebru. "An Efficient Satellite Images Classification Approach Based on Fuzzy Cognitive Map Integration with Deep Learning Models Using Improved Loss Function." IEEE Access, 2024.

[6]. Wang, Ruikun, Lei Ma, Guangjun He, Brian Alan Johnson, Ziyun Yan, Ming Chang, and Ying Liang. "Transformers for Remote Sensing: A Systematic Review and Analysis." Sensors Vol. 24, no. 11, 2024.

[7]. Wang, Keyan, Feiyu Bai, Jiaojiao Li, Yajing Liu, and Yunsong Li. "MashFormer: A novel multiscale aware hybrid detector for remote sensing object detection." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 16, 2023.

[8]. Lu, Wang, Yasheng Zhang, Canbin Yin, Caiyong Lin, Can Xu, and Xin Zhang. "A deformation robust ISAR image satellite target recognition method based on PT-CCNN." IEEE Access, Vol. 9 pp. 23432-23453, 2021

[9]. Alshehri, Mariam, Anes Ouadou, and Grant J. Scott. "Deep Transformer-based Network Deforestation Detection in the Brazilian Amazon Using Sentinel-2 Imagery." IEEE Geoscience and Remote Sensing Letters, 2024.

[10]. Xu, Zhiyong, W Zhang, T Zhang, Zhifang Y, and Jiangyun Li. "Efficient transformer for remote sensing image segmentation." Remote Sensing, Vol. 13, no. 18, 2021.

[11]. Wu, Jiahao, Yongkai Zhao, Ruihan Zhang, Xin Li, and Yuxin Wu. "Application of three Transformer neural networks for short-term photovoltaic power prediction: A case study." Solar Compass, Vol. 12, 2024.

[12]. Lee, Youngchan, and Wonsang You. "Ebat: Enhanced bidirectional and autoregressive transformers for removing hairs in hairy dermoscopic images." IEEE Access, Vol.11 pp.14225-14235, 2023.

[13]. Chen, Ziyang, Wenting Li, Zhongwei Cui, and Yongjun Zhang. "Surface Depth Estimation from Multi-view Stereo Satellite Images with Distribution Contrast Network." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.

[14]. Wensel, James, Hayat Ullah, and Arslan Munir. "Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos." IEEE Access, 2023.

[15]. Shan, Chuanhui, Xinlong Geng, and Chao Han. "Remote sensing image road network detection based on channel attention mechanism." Heliyon 10, no. 18, 2024.

[16]. Z. Yin and A.S. Leon, “Riverine flood hazard map prediction by neural networks”, HydroResearch, DOI. https://doi.org/10.1016/j.hydres.2024.10.003, 2023.

[17]. Verdone, Alessio, Simone Scardapane, and Massimo Panella. "Explainable Spatio-Temporal Graph Neural Networks for multi-site photovoltaic energy production." Applied Energy, Vol. 353, 2024.

[18]. Peña, Francisco J., C Hübinger, Amir H. Payberah, and F Jaramillo. "DeepAqua: Semantic segmentation of wetland water surfaces with SAR imagery using DNN without manually annotated data" International Journal of Applied Earth Observation and Geoinformation, Vol. 126, 2024.

[19]. Bui, Duc Viet, Masao Kubo, and Hiroshi Sato. "Cross-view geo-localization for autonomous UAV using locally-aware transformer-based network." IEEE Access, 2023.

[20]. Ma, Boyi, Falin Wu, Tianyang Hu, Loghman Fathollahi, Xiaohong Sui, Yushuang Liu, and Byambakhuu Gantumur. "Label-driven graph

convolutional network for multi-label remote sensing image classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023.

[21]. Shao, Yanda, Ling Li, Jun Li, Qilin Li, Senjian An, and Hong Hao. "3D displacement measurement using a single-camera and mesh deformation neural network." Engineering Structures, Vol. 318, 2024.

[22]. Zhuang, Jiedong, Xuruoyan Chen, Ming Dai, Wenbo Lan, Yongheng Cai, and Enhui Zheng. "A semantic guidance and transformer-based matching method for UAVs and satellite images for UAV geo-localization." IEEE Access, Vol. 10, pp. 34277-34287, 2022.

[23]. Liu, Xuanguang, Chenguang Dai, Zhenchao Zhang, Mengmeng Li, Hanyun Wang, Hongliang Ji, and Yujie Li. "TBSCD-Net: A Siamese Multi-task Network Integrating Transformers and Boundary Regularization for Semantic Change Detection from VHR Satellite Images." IEEE Geoscience and Remote Sensing Letters, 2024.

[24]. Kaselimi, Maria, Athanasios Voulodimos, Ioannis Daskalopoulos, Nikolaos Doulamis, and Anastasios Doulamis. "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring." IEEE Transactions on Neural Networks and Learning Systems, Vol. 34, no. 7,pp. 3299-3307, 2022.

[25]. Lee, Matthew Chung Hai, Kersten Petersen, Nick Pawlowski, Ben Glocker, and Michiel Schaap. "TETRIS: Template transformer networks for image segmentation with shape priors." IEEE transactions on medical imaging, Vol. 38, no. 11, pp. 2596-2606, 2019.

[26]. Semenov, Alexander, Maciej Rysz, and Garrett Demeyer. "Deep Learning Approach for SAR Image Retrieval for Reliable Positioning in GPS-Challenged Environments." IEEE Transactions on Geoscience and Remote Sensing, 2024.

[27]. Zhao, Jinling, Hao Yan, and Linsheng Huang. "A joint method of spatial–spectral features and BP neural network for hyperspectral image classification." The Egyptian Journal of Remote Sensing and Space Science, Vol. 26, no. 1, pp. 107-115, 2023.

[28]. Rombado, Lucian, Marko Orescanin, and Mara Orescanin. "Uncertainty-Aware Aerial Coastal Imagery Pattern Recognition Through Transfer Learning with ImageNet-1K Variational Embeddings." IEEE Access, 2024.

[29]. Boulila, Wadii, Hamza Ghandorh, Sharjeel Masood, Ayyub Alzahem, Anis Koubaa, Fawad Ahmed, Zahid Khan, and Jawad Ahmad. "A transformer-based approach empowered by a self-attention technique for semantic segmentation in remote sensing." Heliyon, Vol. 10, no. 8, 2024.

[30]. Aleissaee, Abdulaziz Amer, Amandeep Kumar, Rao Muhammad Anwer, Salman Khan, Hisham Cholakkal, Gui-Song Xia, and Fahad Shahbaz Khan. "Transformers in remote sensing: A survey." Remote Sensing, Vol. 15, no. 7, 2023.

[31]. Shafaey, Mayar A., Mohammed A-M. Salem, Hala Mousher Ebied, Maryam N. Al-Berry, and Mohamed F. Tolba. "Deep learning for satellite image classification." In International Conference on Advanced Intelligent Systems and Informatics, pp. 383-391, 2018.

[32]. Alzahem, Ayyub, Wadii Boulila, Anis Koubaa, Zahid Khan, and Ibrahim Alturki. "Improving satellite image classification accuracy using GAN-based data augmentation and vision transformers." Earth Science Informatics,Vol. 16, no. 4, pp. 4169-4186, 2023.

[33]. Yaloveha, Vladyslav, Andrii Podorozhniak, and Heorhii Kuchuk. "Convolutional neural network hyperparameter optimization applied to land cover classification." Radioelectronic and computer systems, Vol. 1, pp. 115-128, 2022.