RESEARCH ARTICLE OPEN ACCESS

An Efficient Deep Learning Model for Fractional Vegetation Index Prediction

Emmanuel Chinyere Echeonwu*, Moses Okechukwu Onyesolu**, Esther Odunayo Oduntan***, Ubom Okon Effiong****

*(Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria
Email: ec.echeonwu@stu.unizik.edu.ng)

** (Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria
Email: mo.onyesolu@unizik.edu.ng)

***(Department of Computer Science , The Federal Polytechnic, Ilaro, Nigeria
Email: odunayo.oduntan@federalpolyilaro.edu.ng)

**** (Department of Geoinformatics & Surveying, University of Uyo, Nigeria,
Email: okonubom@uniuyo.edu.ng)

--************************----------------------------------

Abstract:

This paper introduces the Efficient Vision Transformer (EVT), a deep learning model designed for precise prediction of the Fractional Vegetation Index (FVI) in vegetation monitoring. In order to overcome the limitations of existing methods, EVT incorporates resilient attention mechanisms and streamlined computations, resulting in high accuracy while utilizing fewer computer resources. Its architecture, which includes patch embedding, Transformer encoder layers, and a linear decoder, enables the rapid processing of local information and the comprehensive creation of global context, leading to accurate Vegetation Index predictions. Its ability to process high-resolution aerial images opens up new avenues for monitoring vegetation health and productivity across vast landscapes, providing crucial insights for sustainable land management practices. With 17 million parameters compactly packed in 155 MB and lightning-fast CPU inference duration of 606 milliseconds, EVT demonstrates exceptional efficiency, providing near-real-time insights for precision agriculture and environmental monitoring. The model's computational efficiency and low resource requirements make it accessible to a wide range of users, from researchers and environmental agencies to agricultural organizations and individual farmers, enabling data-driven decision-making on a broader scale. EVT's exceptional accuracy is demonstrated by evaluation measures (Mean Absolute Error of 0.03823, Root Mean Square Error of 0.04582, R-squared value of 0.98556), confirming its reliability across various assessment frameworks and geographical scales. The EVT model represents a groundbreaking solution that offers a future where precision and efficiency come together to support informed decision-making for ecosystem health and sustainable land management.

Keywords **—Fractional Vegetation Index, Attention Mechanisms, Remote Sensing, Precision Agriculture, Environmental Monitoring, Deep Learning, Optimization**

--************************----------------------------------

I. INTRODUCTION

A vegetation map is a two-dimensional graphical representation of a location-based vegetation monitoring system [6]. The vegetation index (VI), which is a highly sought-after source of information for informed conservation decisions and data on biodiversity and natural resources, summarizes this information within a spatial framework. Fractional vegetation cover (FVC), the ratio of the area occupied by vegetation on the ground to the total vegetation area, is another term used to refer to the proportion of vegetative area to the overall area of interest [13]. FVC is a significant biophysical metric that describes the Earth's surface system and is considered essential for investigating the interactions between the aerosphere, pedosphere, hydrosphere, and biosphere [8]. The importance of monitoring this parameter was emphasized in Zeng, X. et al.,[18] as it provides a deeper understanding of land-surface processes, climate change, and numerical weather forecasting [18].

Two broad methodologies, field and remote sensing techniques, have been developed to determine fractional vegetation cover. However, these methods require substantial human effort, expertise, and supervision and do not provide real-time data. The variations in regional climate and temperature can result in changes in vegetation to better adapt to these differences. Consequently, the vegetation zones in a country demonstrate the relationship between vegetation and climatic conditions in different geographic regions across the nation. Nigeria, with its diverse climatic conditions, experiences a significant influence on the numerous plant zones within the country [11]. Temperature variations, weather conditions, and other geographical factors contribute to the variations in vegetation from the northern to southern hemispheres of Nigeria. Each species in Nigeria's vegetation exhibits distinct characteristics that distinguish it from others, resulting in a highly diverse vegetation landscape.

Mapping vegetation is a crucial aspect of carbon cycle studies. Empirical approaches collect sufficient and accurate data from a large number of samples to establish statistical correlations between FVC and vegetation indicators or individual band reflectance. Empirical techniques can achieve appropriate accuracy when applied to specific vegetation types and at a regional scale. However, these methods become less effective in large-scale regions due to uncertainties resulting from different vegetation types and land conditions, making empirical methods inadequate [2].

Spectral Mixture Analysis (SMA) is a technique that decodes the spectral information in remote sensing images (including green vegetation, senescent vegetation, various soil types, and water) by assuming that spectral variation is generated by a small number of surface materials [10]. SMA is based on the assumption that the spectral signature of a pixel is a linear combination of the endmember spectra [9], where each endmember represents a distinct spectral signal that differentiates it from other surface materials or land cover types on the planet. It is commonly assumed that the spectral diversity within an endmember is minimal or nonexistent. Selecting appropriate endmembers is crucial for the success of a mixing model [3]. These endmember signatures can either be directly chosen from an image (image endmembers) or derived from field or laboratory spectra of known materials [1]. However, the primary limitation of pixel unmixing models lies in estimating representative endmembers, as this task is challenging due to the complex land surface conditions and diverse spectral features found at a large scale [20].

In recent years, statistical machine learning approaches have gained popularity in the process of retrieving FVC values. These approaches are computationally efficient and consistently perform well in non-linear fitting [13]. Three main methods exist for esimating the FVC: empirical methods, pixel unmixing modelling and physical model-based methods [15]. Machine learning methods estimate FVC by training on a representative sample database containing pre-processed reflectance and related simulated land surface parameter data, which occurs during the learning phase [13]. Several algorithms and machine learning methods for generating FVC products on regional and global scales have been presented by Jia, K. et al. These strategies have shown satisfactory results, although some effort in feature engineering is required [7].

In this paper, we propose the EfficientVisionTransformer (EVT) for improved computation and accuracy. The EVT focuses on image regression with a specific emphasis on forecasting the critical Fractional Vegetation Index (FVI). By employing robust attention mechanisms and deliberate optimizations, the EVT achieves remarkable accuracy in FVI estimation while requiring fewer computational resources. This enables broader access to various vegetation monitoring scenarios, ranging from rural areas to global satellite data. The capabilities of EVT accelerate research and provide actionable information for precision agriculture, environmental monitoring, and sustainable land management, thereby paving the way for informed decisions that support thriving ecosystems.

II. RELATED WORKS

Remote sensing has a wide range of applications in the implementation of software systems. These applications include the analysis of agricultural diseases, the classification of different types of seeds, and object recognition. In this section, we provide a review of related studies on software models used in remote sensing for vegetation monitoring and optimization efforts.

In a study conducted by Zhang L. et al.[19], the deep learning ENVINet-5 model was employed to monitor vegetation coverage in a research area using medium-resolution Landsat Thematic Mapper (TM) and Operational Land Imager (OLI) satellite images. The training and verification samples were manually labeled using a fusion image and a high-resolution satellite image obtained from Google Earth. The labels were classified into four types of ground objects: desert, water body, cultivated land, and construction land. The normalized difference vegetation index (NDVI) was calculated for the area of interest using this data [19].

Yu, R. et al. [16] proposed an innovative approach to estimate fractional vegetation cover (FVC) using deep transfer learning. The method consisted of two steps. In the first step, a physical model known as the PROSPECT + SAIL radiative transfer model (PROSAIL) was utilized to generate a large number of simulated training samples. In the second step, a long short-term memory network (LSTM) was pretrained using the simulated training dataset obtained in the first phase. The pretrained network was then fine-tuned using a small number of real samples derived from satellite images.

Zeebaree, S. R. et al. [17] introduced the Convolutional Stacked AutoEncoder Recurrent Neural Network (CSAERNet) as an efficient deep learning architecture for classification. This involved utilizing a convolutional neural network (CNN) to extract features from an input, feeding these features into a stacked autoencoder (SAE) for dimensionality reduction, and finally passing the reduced dimensions into a recurrent neural network (RNN) to enhance accuracy [17].

Fan, S. et al. [5] suggested a plant recognition approach based on a deep fully convolutional network with feature fusion (FCN-FF). Initially, the FCN-FF extracted hierarchical characteristics from optical remote sensing images and then fused high-level and low-level information from a deep layer and a shallow layer to identify vegetation. The fusion of multiple layers of features improved the accuracy of vegetation recognition in complex images [5].

Sadeghi-Tehran, P. et al [12] developed a multi-feature learning approach for measuring vegetation development in outdoor field situations. The newly proposed method was compared with state-of-the-art and earlier learning approaches for digital images. The methods were evaluated under various environmental conditions using criteria such as (1) comparison with ground-truth photos, (2) variance over the course of a day due to variations in ambient lighting, (3) comparison with manual measurements, and (4) evaluation of performance over the course of a wheat canopy's life cycle. The authors claimed that the strategy successfully addressed environmental challenges encountered in field conditions [12].

Traditionally, convolutional neural networks (CNNs) excel in local feature extraction but struggle to capture long-range correlations within images. Transformers' self-attention processes bridge this gap, providing a comprehensive understanding of vegetation patterns and leading to more accurate outcomes. Several studies suggest that Transformer-based models outperform traditional CNNs in fractional vegetation index (FVI) prediction accuracy. For instance, Efficient ViT achieves cutting-edge performance on remote sensing datasets, indicating its potential for high-precision vegetation assessments. Transformers are capable of processing not only FVI values but also multiple vegetation index (VI) values simultaneously, providing a more complete picture of vegetation health and environmental conditions. This expands the range of possible applications for VI prediction tasks.

The development of Vision Transformers (ViTs), a unique design that exploits the power of attention processes to capture long-range dependencies within images, has resulted in a paradigm change in computer vision research. Dosovitskiy et al. first developed ViTs, which have exhibited extraordinary performance in a variety of visual comprehension tasks, challenging convolutional neural networks' (CNNs) long-standing supremacy[4].

The attention mechanism is central to ViTs, allowing the model to selectively focus on significant sections of the input image while taking into account their interactions with other regions. This is accomplished by the self-attention operation, which computes the pairwise similarities of all input pieces, allowing the model to capture complex spatial and semantic linkages [14].

The ViT attention mechanism is inspired by the Transformer architecture, which was first developed for natural language processing tasks [14]. Images, on the other hand, are two-dimensional entities by definition, as opposed to sequential data in words. To adapt Transformers to the visual domain, ViTs divide the input picture into non-overlapping patches, which are then linearized and fed sequentially into the Transformer encoder [4].

III. METHODOLOGY

The model is made up of the Patch embedding layer, the encoder layer, and the linear layer. In order to accurately predict vegetation indices (VIs), the EfficientVisionTransformer utilizes a well-coordinated pipeline. The first stage, patch embedding, divides the input image into smaller patches, making them more manageable for further processing. Each patch is then projected into a high-dimensional vector that captures its main features. This stage, similar to linear convolutional filters collecting local information, lays the foundation for subsequent processing.

After patch embedding, the model rearranges these vectors into a sequence compatible with the Transformer encoder. This is where the power of multi-head attention comes into play, allowing the model to attend to multiple sections of the image simultaneously. The model dynamically establishes and re-establishes associations between different patches, just like a language model deriving context from words in a sentence. Through these complex attention mechanisms, the model captures intricate long-range relationships within the image, which are crucial for interpreting subtle vegetation patterns.

The Transformer encoder incorporates feed-forward layers, which introduce non-linearity and expressiveness to these dependencies. Finally, by averaging the attention outputs across patches, a global context is generated, resulting in a single vector that summarizes the entire image. This vector is then fed into the final layer, a linear decoder, which predicts the VI value. Two notable enhancements used in the model are sandwich encoder layers and cascaded group attention, which greatly improve efficiency. These strategies effectively coordinate information flow and attention mechanisms, enabling the model to achieve state-of-the-art accuracy while remaining computationally feasible.

In summary, the EfficientVisionTransformer integrates multiple stages, ranging from local feature extraction to global context building, to produce precise and efficient VI predictions. This paves the way for novel applications in vegetation monitoring and ecosystem management. To summarize the key points:

3.1. Patch Embedding Layer

1) Division: Divides the input image into 16x16 pixel patches (3 channels for RGB).
2) Embedding: Utilizes a convolutional layer to embed each patch into a 600-dimensional vector.
3) Spatial Preservation: Maintains patch placements to improve spatial awareness.

3.2. Transformer Encoder

Encoder Layers: The Transformer consists of six encoder layers.

Multi-Head Attention: Each layer employs multi-head attention with 8 heads, allowing for parallel attention to different regions of the image for deeper understanding.

Feed-Forward Network: Each layer includes a feed-forward network with 2048 hidden units, introducing non-linearity and greater expressiveness.

3.3. Linear Decoder

Encoder Layers: The Transformer consists of six encoder layers.The training procedure of the EfficientVisionTransformer aims to minimize the Mean Squared Error (MSE) between its predicted vegetation indices (VI) and the actual values for aerial photographs. To achieve this, a dataset consisting of 3700 training photos, 284 validation images, and 16 test images is utilized. Each image is in the RGB format and has dimensions of 600x600, encoding valuable data on vegetation.

The primary training loop operates over multiple epochs and generates VI predictions through a forward pass on the training images. The MSE loss between these predictions and the ground truth labels guides the model's improvement via a backward pass and weight updates facilitated by the Adam optimizer. The training and validation losses are continuously monitored during the procedure, providing insights into the learning progress of the model.

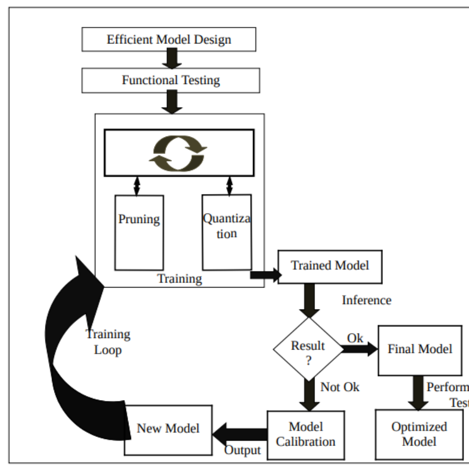

Fig. 1 Efficient model development diagram

Two key optimization strategies, pruning and quantization, are employed to enhance the efficiency and deployment potential of the model. The pruning strategy systematically removes insignificant weights from the model, reducing its size and mitigating the risk of overfitting. Notably, starting from the fourth epoch, a 40% pruning rate was implemented. On the other hand, quantization involves converting the 32-bit floating-point weights to 8-bit integers. This conversion significantly reduces the model's memory footprint and potentially offers computational advantages during deployment.

In addition to these optimization techniques, the training procedure incorporates an accurate loss function and effective data calibration strategies to further improve the EfficientVisionTransformer. This includes scaling the input data using a scalar function before training and using an inverse transformation function during inference to calibrate on the test data. This prior information aids in iterative model training, enabling precise VI predictions and facilitating the model's suitability for efficient real-world applications.

TABLE I
CALIBRATION SET (EPOCH = 5)

S/N	Training	Calibration set
1	1	Min = 0 (initialization set) Max = 1
2	2	Min = 1 Max = 6
3	3	Min = 1.8 Max = 7.10

TABLE II
PRUNING AND QUANTIZATION SCHEDULE

Epoch	Pruning	Quantization
1	0	int8
2	None	None
3	None	None
4	40	Int8
5	None	None

IV. RESULTS AND DISCUSSION

With its complexity of 17 million parameters, the model possesses the necessary intricacy to effectively learn complex vegetation patterns, a crucial factor in ensuring accurate vegetation index (VI) prediction. It manages to strike a practical balance by maintaining a relatively compact size of only 155 MB, making it well-suited for implementation on devices with limited memory capacity. The model's efficiency is exemplified by its remarkably fast CPU inference time of 606 milliseconds, allowing for near-real-time VI insights. This capability opens up possibilities for various applications, such as precision agriculture and environmental monitoring. Furthermore, the model operates seamlessly with a memory footprint of 362 MB, ensuring uninterrupted performance. The EfficientVisionTransformer presents a compelling combination of accuracy and efficiency, presenting numerous opportunities in the field of VI prediction, ranging from agricultural management to environmental research.

The model demonstrates exceptional accuracy, as demonstrated by the following metrics:

TABLE III
DESCRIPTIVE METRICS

S/N	Metrics	Value
1	Mean Absolute Error	0.03823
2	Mean Square Error	0.00299
3	Root Mean Square Error	0.04582
4	R-Squared	0.98556

A. Low average errors: The model exhibits a Mean Absolute Error (MAE) of 0.03823 and a Root Mean Square Error (RMSE) of 0.04582. These values indicate that the model's predictions closely align with the true values, with minimal variances.

B. Strong goodness-of-fit: The high R-squared value of 0.98556 signifies a significant correspondence between the model's predictions and the actual data. This suggests that the model successfully captures the underlying patterns and relationships.

C. Adaptability across scales: The consistency of the MAE and RMSE values showcases the model's accuracy regardless of the scale used for error measurement. This substantiates the reliability of the model across various evaluation frameworks.

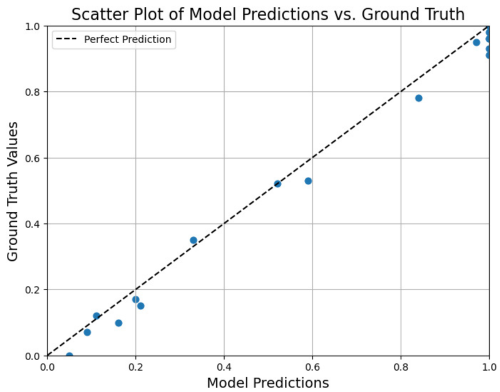

Fig. 2Scatter Plot

The scatter plot representing predictions versus ground truth exhibits a generally positive linear correlation, suggesting that the model successfully captures certain inherent data patterns. Nevertheless, there are notable deviations from the optimal 45-degree line, signifying potential areas for further improvement. Meanwhile, a histogram illustrating prediction errors displays a reasonably normal distribution centered around zero, which is typically desirable. However, the small sample size poses challenges in corroborating this normality.

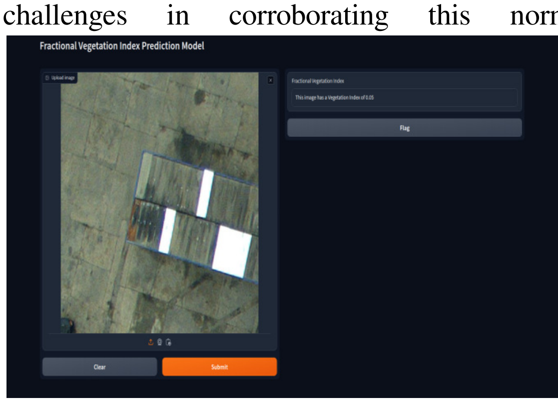

Fig. 3 Model output

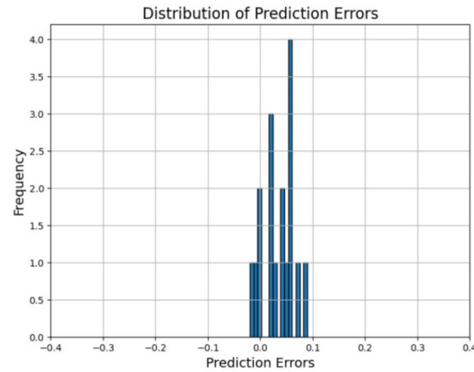

Fig. 4 Distribution of Prediction Errors

V. CONCLUSIONS

The study presents the EfficientVisionTransformer (EVT) as a novel solution for predicting the Fractional Vegetation Index (FVI) in order to overcome limitations in current vegetation monitoring systems. By developing EVT, which incorporates robust attention mechanisms and streamlined calculations, the study addresses the issues associated with resource-intensive and less effective approaches. This advancement allows EVT to achieve exceptional accuracy in FVI prediction while utilizing fewer processing resources. EVT efficiently processes local information and establishes global context by integrating patch embedding, Transformer encoder layers, and a linear decoder, resulting in accurate and rapid VI predictions. Future research opportunities include further enhancing the EVT model to improve prediction accuracy and reduce processing requirements for widespread implementation. Exploring applications beyond vegetation monitoring, such as climate change analysis and urban development planning, could enhance the usefulness of EVT. Additionally, expanding datasets and developing methodologies for real-time applications would enhance the model's ability to generalize and be practical in supporting informed decisions for ecosystem health, sustainable land management, and environmental conservation. In conclusion, the Efficient Vision Transformer represents a significant advancement in vegetation monitoring, combining precision, efficiency, and scalability to enable precise ecosystem assessments and informed decision-making.

ABBREVIATIONS

CNN: Convolutional Neural Network
CPU: Central Processing Unit
EVT: EfficientVisionTransformer
FVC: Fractional Vegetation Cover
FVI: Fractional Vegetation Index
MAE: Mean Absolute Error
MB: Megabyte
MSE: Mean Square Error
NDVI: Normalized Difference Vegetation Index
OLI: Operational Land Imager
RGB: Red Green Blue
RMSE: Root Mean Square Error
SAE: Stacked AutoEncoder
SMA: Spectral Mixture Analysis
TM: Thematic Mapper
VI: Vegetation Index
ViT: Vision Transformer

ACKNOWLEDGMENT

I am grateful to Google for supporting this project through the EMEA PhD Research Fellowship. This expedition would have been significantly more difficult, if not impossible, without your generosity. I also want to thank my superivor, Prof M.O Onyesolu for his support and encouragement.

REFERENCES

[1] Adams, J.B., Smith, M.O & Johnson, P.E. (1986). Spectral mixture modeling: A new analysis of rock and soil types at the viking lander 1 site. *Journal of Geophysics 91*, 8098–8112. https://doi.org/10.1029/JB091iB08p08098

[2] Asner, G.P & Heidebrecht, K.B. (2002). Spectral unmixing of vegetation, soil and dry carbon cover in arid regions: Comparing multispectral and hyperspectral observations. *International Journal in Remote Sensing. 23*, 3939–3958. https://doi.org/10.1080/01431160110115960.

[3] Deng, Z., Lu, Z., Wang, G., Wang, D., Ding, Z., Zhao, H., Xu, H., Shi, Y., Cheng, Z & Zhao, X. (2021). Extraction of fractional vegetation cover in arid desert area based on ChineseGF-6 satellite. *Open Geosciences*, *13*(1), 416-430. https://doi.org/10.1515/geo-2020-0241

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T.,... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[5] Fan, S., Li, Y., Yan, Z., Guo, L & Wang, X. (2017). *Vegetation Recognition Based on Deep Learning with Feature Fusion. 19-23.* https://doi.org/*10.1145/3133264.3133276*

[6] Janssen, J. A. M. 2004. *Community Ecology.* The use of sequential vegetation maps for monitoring in coastal areas. *5*(1), 31–43. .https://doi.org/10.1556/comec.5.2004.1.4

[7] Jia, K., Liang, S., Liu, S., Li, Y., Xiao, Z., Yao, Y., Jiang, B., Zhao, X., Wang, X & Xu, S. (2015). Global land surface fractional vegetation cover estimation using general regression neural networks from MODIS surface reflectance. *IEEE Transaction. Geoscience. Remote Sensing. 53*, 4787–4796. https://doi.org/10.1109/TGRS.2015.2409563.

[8] Jia, K., Li, Y., Liang, S., Wei, X., & Yao, Y. (2017). Combining estimation of green vegetation fraction in an arid region from Landsat 7 ETM+ Data. *Remote Sensing*, *9*(11), 1121. https://doi.org/10.3390/rs9111121.

[9] Malhi, R., Srivastava P & Kiran, G. (2020). Identification of functionally distinct plants using linear spectral mixture analysis. In *Hyperspectral Remote Sensing: Theory and Applications*. Elsevier. https://doi.org/10.1016/B978-0-08-102894-0.00008-5.

[10] Manolakis, D., Lockwood, R. & Cooley, T. (2016). Spectral Mixture Analysis. In *Hyperspectral Imaging Remote Sensing: Physics, Sensors, and Algorithms*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316017876.010.

[11] Mfonobong, D.(2020). Vegetation zones in Nigeria and their characteristics. Available from: *https://nigerianinfopedia.com.ng/vegetation-zones-in-nigeria/* [Accessed 10 March 2022].

[12] Sadeghi-Tehran, P., Virlet, N., Sabermanesh, K. & Hawkesford, M. J. (2017). Multi-feature machine learning model for automatic segmentation of green fractional vegetation cover for high-throughput field phenotyping. *Plant Methods.13* (103)1-16. https://doi.org/10.1186/s13007-017- 0253-8.

[13] Song, W., Tian, Z., Xihan, M., Bo, Z., Jing, Z., Guangjian, Y., Li, W & Zheng, N. 2022. Using a vegetation index-based mixture model to estimate fractional vegetation cover products by jointly using multiple satellite data: method and feasibility analysis. Forests 13(5): 691. https://doi.org/10.3390/f13050691.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://arxiv.org/abs/1706.03762.

[15] Yang, L., Jia, K., Liang, S., Liu, J & Wang, X. (2016). Comparison of four machine learning methods for generating the GLASS fractional vegetation cover product from MODIS data. *Remote Sensing. 8*:682. https://doi.org/10.3390/rs8080682.

[16] Yu, R., Li, S., Zhang, B & Zhang, H. (2022). A deep transfer learning method for estimating fractional vegetation cover of sentinel-2 multispectral images. IEEE Geoscience and Remote Sensing Letters, 19: 1-5. https://doi.org/10.1109/LGRS.2021.3125429.

[17] Zeebaree, S.R., Ahmed, O & Obid, K. (2020).CSAERNet: An efficient deep learning architecture for image classification. 2020 3rd International Conference on Engineering Technology and its Applications (IICETA). 122-127. https://doi.org/10.1109/IICETA50496.2020.9318859.

[18] Zeng, X.B., Dickinson, R.,Walker, A., Shaikh, M., Defries, R & Qi, J. (2000). Derivation and evaluation of global 1-km fractional vegetation cover data for land modeling. *Journal of Applied Meteorology. 39*, 826-839.https://doi.org/10.1175/1520-0450(2000)039<0826:DAEOGK>2.0.CO;2.

[19] Zhang L., Wang, W., Dong, Y & Wu, L. (2022).Vegetation coverage monitoring model design based on deep learning. *ScientificProgramming*. https://doi.org/10.1155/2022/4818985.

[20] Zou, J., Lan, J & Shao, Y. (2018). A hierarchical sparsity unmixing method to address endmember variability in hyperspectral image. *Remote Sensing*. 10:738. https://doi.org/10.3390/rs10050738.

