Journal of Information Systems Engineering and Management
2025, 10(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Overview of Vision Transformers and Deep Learning Methods for Classifying Remote Sensing Images

Keerthishree P V[1], Suhas G K[1*] ,S G Gollagi[2] , Yathisha L[3]

[1]*Research scholar, Computer Science and Engineering, Akshaya Institute of Technology, Affiliated to Visv esvaraya Technological University Karnataka-572106, India.*
Email: keerthishree53@gmail.com

[1*]*Professor, Research Centre for Computer and Information Sciences, Akshaya Institute of Technology, Affiliated to Visvesvaraya Technological University Karnataka-572106, India.*
Email: Suhask300@gmail.com

[2]*Professor, Computer Science and Engineering, S G Balekundri Institute of Technology, Affiliated to Visvesvaraya Technological University Karnataka, India.*
Email:shantesh1973@rediffmail.com

[3]*Professor, Electronics and Communication Engineering, Akshaya Institute of Technology Affiliated to Visvesvaraya Technological University, Tumkur, India-572106*
Email: yathisha.171@gmail.com

ARTICLE INFO

Received: 25 Sept 2024

Revised: 29 Nov 2024

Accepted: 08 Dec 2024

ABSTRACT

The diversified, multifarious, and high-dimensional nature of remote-sensing photos makes remote-sensing image scene categorization (RSISC) an important and challenging challenge for understanding changes on Earth's surface. RSISC's main goal is to give acquired images semantic labels so that they can be arranged according to semantic content. Deep learning frameworks, especially in image analysis, have seen a sharp increase in interest and development in recent years. Even though deep learning approaches are more computationally costly than conventional machine learning techniques, they have demonstrated great potential in this field. This study provides a thorough evaluation of several deep learning (DL) methods, including Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) like ResNet, VGG16, InceptionV3, and DenseNet. We use the NWPU-RESISC45 and RSI-CB256 remote sensing datasets, which are both publically accessible, to assess how well these models perform. The findings show that although conventional CNN designs perform competitively, Vision Transformers (ViTs) are better at identifying intricate spatial correlations in the data for the categorization of remote sensing images. Because vision transformers use self-attention mechanisms to efficiently capture complicated spatial linkages and long-range dependencies, they perform exceptionally well in remote sensing picture classification. Furthermore, multi-scale feature extraction is made possible by their patch-based processing, which improves accuracy, particularly in high-resolution images.

Keywords: Image classification, vision transformer, deep learning, machine learning, CNN, SVM, VGG, XGBoost, KNN, Random Forest.

I. INTRODUCTION

In many applications, such as monitoring the environment and natural hazards, urban planning and development, item detection, and vegetation mapping, remote sensing image scene classification (RSISC) is a vital and essential task for precise analysis. RSISC systems have undergone a revolution thanks to developments in deep learning models, which have outperformed more conventional methods that depend on machine learning and image processing. In order to further improve RSISC, this work makes use of vision transformer and deep learning technologies. Image categorization and object/target recognition in a range of images, such as medical, thermal, infrared, and remotely sensed images, are among the many uses for deep learning and computer vision [1–4]. Satellite imagery stands out as a primary source for gathering geographic information [5], with numerous urban

infrastructure planning and development applications. The data acquired from satellite sources are immense and growing exponentially. Coping with this volume demands efficient techniques for data extraction is needed. Satellite image classification constitutes a multistage process, commencing with feature extraction from images and culminating in theircategorization [6].Additionally, Chen [7] presented Extreme Gradient Boosting Trees (XGBoost), an improved gradient-boosting technique. A CNN and XGBoost-based image classification algorithm was presented by Ren et al. [8]. XGBoost serves as a recognizer, producing more accurate output, while CNN pulls features from the input. DRSNet, a modified CNN designed for Landsat 8 remote sensing employing small patch sizes, was created by Chen et al. [9]. In order to improve feature extraction, this architecture combines Inception-ResNet with channel attention and includes a unique residual inception channel attention block. Reduction modules take the place of pooling layers to overcome representational constraints. Furthermore, retrieving information lost during earlier down-sampling stages is facilitated by the deliberate employment of up-sampling steps before to final pooling layers. Aggregated Features from Dual Paths (AFDP) was proposed by Shaheed et al. [10] using simplified convolutional neural networks for different image representations. For efficient learning of discriminative image features, the method combines bilinear pooling and feature connection principles with a dual-branch feature extractor with fewer parameters and a novel feature fusion strategy. For remote sensing, Wang et al. [11] suggested a CNN with residual dense attention blocks that prioritizes local data and channel-based multi-instance pooling.For computer vision applications, such as the classification of remote sensing images, deep learning has become a potent tool. The shortcomings of conventional shallow models like Support Vector Machines (SVM) and Random Forests (RF) have been outperformed by its deep architecture and capacity to learn intricate features [13]. In remote sensing applications such segmentation, object detection [17], and classification [18], deep convolutional neural networks (CNN) [16], deep belief networks (DBN) [14], and stacked auto-encoders (SAE) [15] have shown notable results.While deep learning has shown promising results in theclassification of remote sensing images, its application to high-resolution imagery has been hindered by limited datasetsand the computational demands of training complex models [19]. Vision transformers, which leverage self-attention mechanisms, offer potential solutions to these challenges. However, their computational complexity, especially for large images, remains a challenge [20].This paper examines various deep-learning approaches and vision transformer for remote sensing image classification. We compare state-of-the-art Convolution Neural Networks (CNNs) and vision transformers, evaluating their performanceon NWPU-RESISC45 and RSI-CB256 datasets.

II. METHODOLOGY

This section comprehensively compares deep learningtechniques and vision transformers for remote sensing image classification.

A. ResNet

The ResNet system utilizes deep residual networks to enhanceclassification performance by addressing the vanishing gradient problem in deeper networks through a residual process. The ResNet architecture [22] applies residual learning to groups of stacked layers. It also incorporates stacked convolutional layers for feature learning and extraction. The ResNet model consists of five blocks, each with the same convolutional layer size, except for the first block, which performs down-sampling. Each block includes a composite function made up of batch normalization (BN), a non-linear transformation unit, a rectified linear unit (ReLU),and a convolution layer. A skip connection is employed to bypass the non-linear transformations using an identity function. Deep features are extracted and down-sampled through integrated pooling units, including Maxpool, AdaptiveAvgPool, and AdaptiveMaxPool.

Mathematical representation is defined as:

$$y = F(x, Wi) + x \quad \text{.....(1)}$$

let the input and output vectors of a layer be represented by xand y, respectively. The function F(x, Wi) represents the residual mapping to be learned through multiple convolutional layers and operators. After this, the feature maps are added element-wise, channel by channel.

The parameters of this model are outlined as follows:

1. Total number of parameters: 44,611,648

2. Trainable parameters: 2,216,832
3. Non-trainable parameters: 42,394,816

An overview of the ResNet50 layered architecture is summarized below:

- Block A: 1 unit, Conv, ReLU, MaxPool, BatchNorm,112 × 112 resolution, 64 channels
- Block B: 15 units, Conv, ReLU, BatchNorm, resolutions: 56 × 56, 28 × 28, 14 × 14, 7 × 7, channels: 256, , 2048
- Block C: 10 units, Conv, BatchNorm, resolutions: 56

  × 56, 28 × 28, 14 × 14, 7 × 7, channels: 64, 128, ,

  2048
- Block D: 12 units, Conv, ReLU, BatchNorm,AdaptiveMaxPool, AdaptiveAvgPool, 7 × 7

  resolution, 2048 channels
- Block E: 1 unit, Linear, ReLU, BatchNorm, 7 × 7resolution, 512 channels
- Block F: 1 unit, Linear, 7 × 7 resolution, 10 channels.

B. VGG16

The VGG architecture [23] enhances the basic ConvNetdesign by progressively increasing the network's depth with additional convolutional layers, leading to a substantial improvement in accuracy. Input images are passed through aseries of convolutional (conv) layers of varying sizes, followed by non-linearity via the ReLU activation function, batch normalization, and pooling units such as average pooling, max pooling, adaptive average pooling, and adaptivemax pooling. These pooling layers are used to maintain spatialresolution after the convolution process and are applied overa 2 × 2-pixel window. The VGG16 model is organized into aseries of blocks. The basic ConvNet uses Eq. (2) for feature extraction.

$$F_i = \text{ReLU}\,(W \times F_{i-1} + b_i) \quad \text{.....(2)}$$

Here, F_i represents the feature map of the current layer, F_{i-1} denotes the feature map from the previous layer, W is the filter kernel, and b_i is the bias added to the feature map of each layer. The rectified linear unit (ReLU) activation function is defined as:

$$U(y) = \max(0, y) = y \text{ if } y \geq 0$$

$$0 \text{ if } y < 0 \quad \text{.....(3)}$$

Where y is the resulting feature map.

The parameters of this model are outlined as follows:

1. Total number of parameters: 11,117,632
2. Total trainable parameters: 532,480
3. Total non-trainable parameters: 10,585,152

C. Inceptionv3

For dependable classification performance, the Inception model uses several Inception layers. Inception modules use layered 1 x 1 convolutions to reduce dimensionality, allowing for deeper networks and more efficient computing [24]. By using a multiscale approach, the Inception-v3 model increases the network's depth and breadth. Deeper structures that improve the classification effectiveness of remote-sensing satellite images are made possible by this design, which also helps to combat vanishing gradient problems. An integrated pooling system is used to extract and down-sample deep features. The following are the model's parameters:

1. Total number of parameters: 23,897,056
2. Trainable parameters: 2,145,920
3. Non-trainable parameters: 21,751,136

D. DenseNet

The goal of DenseNets, which were first presented in [21], is to take advantage of feature reuse inside the network to produce small, extremely parameter-efficient models that are simple to train. The network is made up of several dense blocks, and before every three × three convolution layer, a 1 × 1 convolution layer is added to increase computational efficiency inside each block. As a result, fewer input feature maps—which are usually more than output feature maps—are produced. The following equation represents dense blocks and the concatenation process:

$$y = C_n([x_0, x_1, \ldots, x_{n-1}]) \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

where x_0, ,x_{n-1} represent the concatenation of input feature

maps from the convolutional operators C_n. Here n denotes the number of blocks with the same structure, F refers to the operators, and H and W represent the resolutions.

The parameters of this model are as follows:

1. Total number of parameters: 8,012,672
2. Trainable parameters: 1,142,464
3. Non-trainable parameters: 6,870,208

E. Vision transformer-based architecture

Through the use of a multi-head attention mechanism, vision transformers effectively learn multi-scale, multi-resolution, and high-level spatial characteristics by extracting both local and global contexts. A global average pooling system is used to concatenate and up-sample the dense feature maps that are generated. In order to efficiently collect complex characteristics in remote sensing satellite images, this method combines global average pooling with local and global attention. As seen in Figure 1, the total procedure includes processes like flattening, tokenization, position embedding, and classification. The Transformer encoder specifically splits the input image into fixed-size patches, flattens them, embeds them linearly, combines them with position embeddings, and then passes them through. The steps for training the vision transformer are shown in Algorithm 1.

The parameters for this model are as follows:

1. Total number of parameters: 4,166,151
2. Trainable parameters: 4,166,151
3. Non-trainable parameters: 0

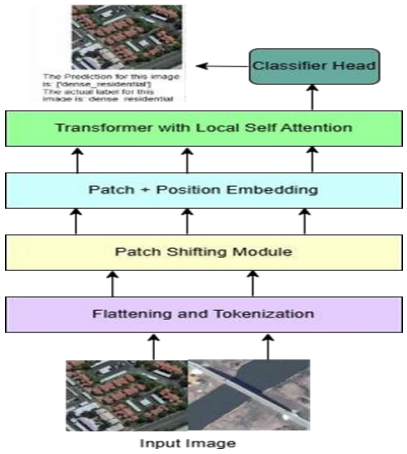

Fig 1. The basic layout diagram for Vision Transformer design

Algorithm 1: Vision TransformerInput: Training images

Output: Predicted labels.

1. B*atchsiz*e is set to 100, the number of iterations is30, image dimensions to 224, Optimizer Adam (learning rate: 0.0003)
2. Set the number of mini-batches.
3. For iteration = 1: Number of iterations
    a. For batch =1: *nb*
    b. Select a batch from the training dataset.,
    c. Create another batch of augmented imagesusing a specific augmentation technique.
    d. Train the model using both the original and augmented images while minimizing the cross-entropy loss.
    e. Perform backpropagation to adjust the loss.
    f. Model parameters are updated.
4. Classify test images.

III. DATASETS

We have experimented with remote sensing picture scene categorization utilizing the RSI-CB256 and NWPU-RESISC45 datasets.

The NWPU-RESISC [12] dataset is created by Northwestern Polytechnical University (NWPU) and serves as a baseline for the categorization of remote sensing images. This dataset consists of 31,500 256 x 256 pixel RGB photos from 45 different scene classes, each with 700 images. Its comprehensiveness, which includes a variety of scene types and a large number of photographs, is one of its noteworthy features. It also includes a broad range of variables, such as viewpoint, illumination, backdrop occlusion, translation, object posture, and spatial resolution. It is noteworthy that every class exhibits significant internal variety while retaining unique traits from other classes. The RSICB256 Satellite Image Classification Dataset, which contains a set of sample images, is also used in this work. This dataset includes snapshots from Google Maps and photos of four different classes that were collected using different sensors. This dataset's visualization highlights how difficult it is. Points of Interest (POIs) from several nations are included, as are remote sensing photos obtained from Google Earth data and Bing Maps, respectively.

IV. RSEULTS AND DISCUSSION

Using standard criteria, we have assessed the performance of the Vision Transformer and five CNN-based deep learning frameworks on two publicly available datasets, RSI-CB256 and NWPU-RESISC45, in this section. The efficacy of each framework in identifying remote sensing images is evaluated using key performance criteria, such as accuracy, precision, recall, and F1-score. The evaluation results of each model on the RSI-CB256 and NWPU-RESISC45 datasets are shown in Tables I and II, respectively. Figures 2 and 3 display the predictions made using Vision Transformer for the NWPU and RSI-CB256 datasets, respectively. Based on the results presented in table I and II vision transformer-based architecture has outperformed otherDL-based architectures. However, the DenseNet framework has also shown excellent performance. While ResNet, Inceptionv3 and VGG16 have average performance. Because of its self-attention mechanism, patch-based processing, and increased flexibility, Vision Transformer-based architectures have special advantages that can improve performance. These advantages make Vision Transformer especially well-suited to capture the complex relationships in remote sensing data, which improves classification accuracy. Images that belong to different classes but share similar properties can be successfully distinguished by the Vision Transformer concept. According to this study, deeper CNNs—like DenseNet121 and ResNet101—perform better than shallower CNNs because they have more convolutional layers. Deeper and more intricate models, however, demand more processing power for both inference and training.A Windows 10 computer with an i5 9500 CPU and 64 GB of RAM was used in this study. The Keras framework is used for deep learning studies. Every network was set up with the same hyperparameters to guarantee authenticity and fairness. Prior to training, pictures were transformed to the RGB color scheme and scaled to 256x256 pixels.

TABLE I COMPARISON OF MODEL PERFORMANCE ON THE RSI-CB256 DATASET.

Methods	Accuracy	Precision	Recall	F1 Score
Vision Transformer	98.4	98	98	98
DenseNet 121	98.1	97	97	97
ResNet101	81.97	81	80	80
Inception V3	75.6	74	73	72
VGG16	75.3	74	74	73

TABLE II COMPARISON OF MODEL PERFORMANCE ON THE RSI-CB256 DATASET.

Methods	Accuracy	Precision	Recall	F1 Score
Vision Transformer	97.7	96	96	96
DenseNet 121	94.4	92	91	91
ResNet101	89.1	87	86	85
Inception V3	75.1	73	73	72
VGG16	74.8	72	72	72

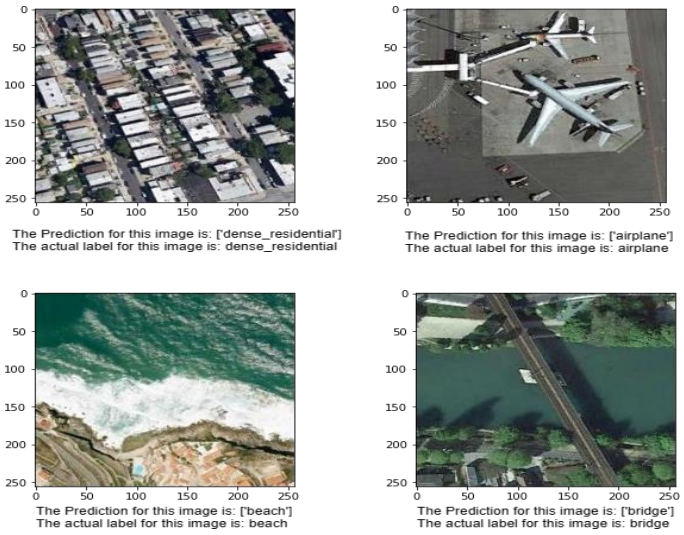

Fig 2. Prediction results of NWPU-RESISC45 dataset.

V. CONCLUSION

In this work, we carried out a comprehensive comparison of several deep learning methods for remote sensing image categorization, including Vision Transformers (ViTs) and conventional Convolutional Neural Networks

(CNNs) like ResNet, VGG16, InceptionV3, and DenseNet. According to our research, Vision Transformers provide significant improvements in accuracy and feature representation, even though CNN architectures have long served as the foundation for image classification tasks. The analysis of several publically accessible datasets showed that ViTs' self-attention mechanisms and patch-based processing methodology enable them to effectively capture intricate spatial linkages and dependencies. These features allow them to perform better in classification, especially when dealing with high-resolution remote sensing photos that contain complex details and a variety of land cover types.However, due to their high memory and processing demands, Vision Transformers may be less effective to train and implement, particularly on smaller datasets. All things considered, this study shows promise for classifying remote sensing data utilizing Vision Transformers and sophisticated deep learning techniques. Future research could look into hybrid models that combine the advantages of ViTs and CNNs, as well as how they might be used in real-time remote sensing situations.

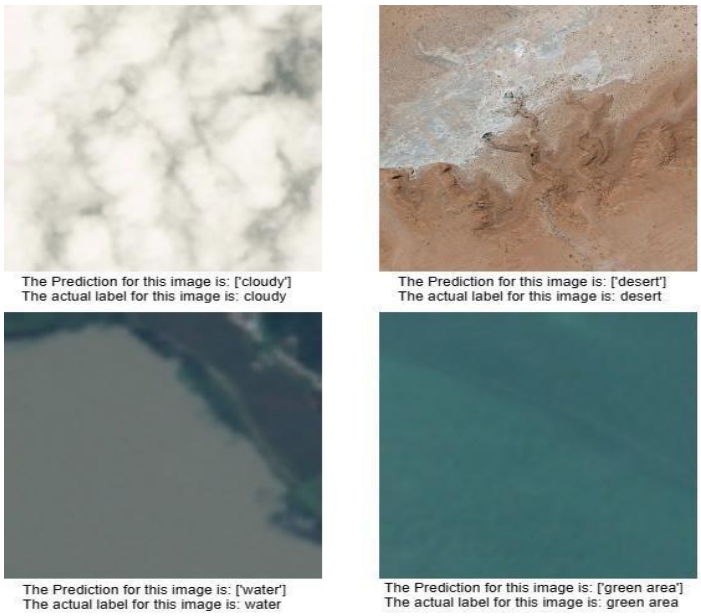

Fig 3. Prediction results of RSI-CB256 dataset.

REFERENCES

[1] C. Shi, X. Zhang, J. Sun, and L. Wang, “Remote sensing scene image classification based on self-compensating convolution neural network,” Remote Sensing, vol. 14, no. 3, p. 545, 2022.

[2] S. Karimi Jafarbigloo and H. Danyali, “Nuclear atypia grading in breast cancer histopathological images based on cnn feature extraction and lstm classification,” CAAI Transactions on Intelligence Technology, vol. 6, no. 4, pp. 426–439, 2021.

[3] X. Zhang and G. Wang, “Stud pose detection based on photometric stereo and lightweight yolov4,” Journal of Artificial Intelligence and Technology, vol. 2, no. 1, pp. 32–37, 2022.

[4] A. Shabbir, A. Rasheed, H. Rasheed et al., “Detection of glaucoma using retinal fundus images: a comprehensive review,” Mathematical Biosciences and Engineering, vol. 18, no. 3, pp. 2033–2076, 2021.

[5] W. Zhang, P. Du, P. Fu et al., “Attention-aware dynamic self-aggregation network for satellite image time series classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–17, 2022.

[6] N. Hamid and J. R. Abdul Hamid, “Multi level image segmentation for urban land cover classifications,” IOP Conference Series: Earth and Environmental Science, vol. 767, no. 1, Article ID 012024, 2021.

[7] [7]. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

[8] [8]. Ren, X.; Guo, H.; Li, S.; Wang, S.; Li, J. A Novel Image Classification Method with Cnn-Xgboost Model; Springer: Cham, Switzerland, 2017.

[9] Chen, Feihao, and Jin Y. Tsou. "DRSNet: Novel Architecture for Small Patch and Low-resolution Remote Sensing Image Scene Classification." International Journal of Applied Earth Observation and Geoinformation, vol. 104, 2021, p. 102577, https://doi.org/10.1016/j.jag.2021.102577. Accessed 23 Dec. 2023.

[10] Shaheed, K., Abbas, Q., Hussain, A., & Qureshi, I. (2022). Optimized Xception Learning Model and XgBoost Classifier for Detection of Multiclass Chest Disease from X-ray Images. Diagnostics, 13(15), 2583. https://doi.org/10.3390/diagnostics13152583.
[11] Wang, Xinyu, et al. "A remote-sensing scene-image classification method based on deep multiple-instance learning with a residual dense attention ConvNet." Remote Sensing 14.20 (2022): 5095.
[12] Cheng, G., Han, J., & Lu, X. (2017). Remote Sensing Image Scene Classification: Benchmark and State of the Art. ArXiv.
[13] Ball JE, Anderson DT, Chan CS. A comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. J Appl Remote Sens. 2017;11(4): 042609.
[14] Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527–54.
[15] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A, Bottou L. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res. 2010;11:12. [16]. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012 (p. 25).
[16] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 (pp. 1–9).
[17] Chen X, Xiang S, Liu C-L, Pan C-H. Vehicle detection in satellite images by hybrid deep convolutional neural networks. IEEE Geosci Remote Sens Lett. 2014;11(10):1797–801.
[18] Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci. 2021;2(6):1–20.
[19] Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. 2021. arXiv preprint arXiv: 2106. 04554
[20] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 (pp. 4700–4708).
[21] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016 (pp. 770–778).
[22] Simonyan K, Zisserman A. Very deep convolutional networks for large- scale image recognition. arXiv preprint arXiv: 1409. 1556 (2014).
[23] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 (pp. 1–9).