Innovative Informatics and Artificial Intelligence Research (IIAIR)
Vol. 1, Issue 1, pp. 12–18, 2025
Received 30 Oct 2024; accepted 23 March 2025; published 30 Apr 2025
https://doi.org/10.35718/iiair.v1i1.1227

Open Plant Species Recognition Using Vision Transformer Network and Top-K Logit Disparity Score

Gusti Ahmad Fanshuri Alfarisy [1], Kassim Kalinaki [2], Owais Ahmed Malik [3], Rizal Kusuma Putra [1], and Aninditya Anggari Nuryono [1]

[1]Department of Informatics, Institut Teknologi Kalimantan, Balikpapan, Indonesia
[2]Department of Computer Science, Islamic University in Uganda, Mbale, Uganda
[3]School of Digital Science, Universiti Brunei Darussalam, Brunei Darussalam

Corresponding author: Gusti Ahmad Fanshuri Alfarisy (gusti.alfarisy@itk.ac.id)

To cite this article: G. A. F. Alfarisy, K. Kalinaki, O. A. Malik, R. K. Putra, A. A. Nuryono, "Open Plant Species Recognition Using Vision Transformer Network and Top-K Logit Disparity Score," *Innovative Informatics and Artificial Intelligence Research*, vol. 1, issue 1, 2025. [Online]. Available: https://doi.org/10.35718/iiair.v1i1.1227

Gusti Ahmad Fanshuri Alfarisy serves as an Editor of IIAIR but was not involved in the peer-review process of this article

Abstract

Reliable plant species identification is essential for biodiversity conservation, agriculture, and ecological research. However, current plant species recognition systems often struggle with the rejection of unknown classes, which limits their applicability in real-world scenarios. Typically, the maximum probability score is used to reject unknown classes, relying solely on the highest output while neglecting the significance of other output scores, which may restrict the model's potential. In this research, we propose a novel scoring function named the Top-K Logit Disparity Score (TKLDS) for open-set plant species recognition using a Vision Transformer (ViT) network. We conducted extensive experiments on the VNPLANT200 dataset consisting of 200 plant species, where the ViT-L/16 model achieved the highest accuracy in closed-set recognition and the highest Area Under the Receiver Operating Characteristic curve (AUROC) between known and unknown classes compared to other state-of-the-art models, such as ResNet, ConvNeXt, Swin Transformer, and MaxViT. Our results indicate that tuning the parameter k in TKLDS consistently improved the arithmetic mean of closed-set accuracy and AUROC across all pre-trained models. Notably, larger values of k generally led to better performance, with the ViT-L/16 model yielding an arithmetic mean score of 0.975 ± 0.005 for $k = 4$ with 5 combinations. These findings demonstrate the potential of TKLDS as a robust scoring function for open-set recognition tasks, highlighting its effectiveness in improving performance metrics in plant species identification.

Keywords: plant species identification; open set recognition; out-of-distribution detection; deep learning; machine learning

1. Introduction

Biodiversity is the epicenter of sustainable living, affecting various aspects of human life, including health, livelihood, agriculture, and other fields such as forestry and biotechnology [1]. In urban areas, ecosystems and biodiversity play a vital role in sustaining urban development by enhancing disaster resilience, improving water and food security, regulating temperature, and providing other benefits [2]. Additionally, Opoku highlights that integrating biodiversity into all development projects presents significant opportunities for the built environment [3]. Marselle et al. suggest that developing biodiversity conservation in urban areas is an investment in public health [4]. Furthermore, biodiversity exposure could enhance the immune system [5]. In Germany, plant species richness contributed positively to mental health [6].

Hence, measuring biodiversity—especially plants—is important, and automatic plant species classification can accelerate the development of biodiversity monitoring systems. This automation can reduce the laborious manual identification process by taxonomists. Traditional identification is very challenging for professionals such as farmers, foresters, conservationists, or landscape architects [7]. Additionally, becoming proficient in the identification of many taxa is difficult [8].

However, existing solutions for automatic species classification using deep learning models still focus primarily on classification performance [9, 10, 11], while lacking the ability to reject unknown classes. Given the vast number of plant species [12], it is crucial to develop robust models capable of identifying classes that were not part of the training set. This would enhance the model's robustness in real-world environments. One way to achieve this is through Open-Set Recognition (OSR), which evaluates performance based on both the correct classification of known classes and the rejection of unknown classes.

Many OSR models, however, rely on maximum probability as the primary score for distinguishing known from unknown classes [13, 14, 15]. While maximum probability indicates the class with the most similar features to a sample, information from other classes could also prove useful in identifying unknown classes. To address this, we investigate the use of top logit information to enhance the performance of unknown class rejection.

This study introduces a deep learning model for rejecting unknown classes in plant species classification, using the Vision Transformer. We propose a novel scoring function, the Top-K Logit Disparity Score (TKLDS), as an alternative to maximum probability or logits, which considers multiple logits to improve the rejection of unknown classes.

The incorporation of multiple logits through TKLDS could impact OSR models across various domains. In biodiversity monitoring, it can enhance model robustness by enabling the rejection of completely unknown classes or serving as an initial step in identifying new species. In agriculture, it can improve the reliability of plant disease recognition or support the collection of new samples for potential emerging diseases. Therefore, we advocate testing this simple approach to enhance the performance of any OSR model.

2. Related Works

In deep learning models, Vision Transformer (ViT) has been employed to identify plant species and has demonstrated outstanding capability. Pan et al. [16] experimented with multimodal features using aerial images and geo-location information. They improved ViT by concatenating the embeddings of both features and using a dynamic transformer encoder that automatically samples the relevant patches and modifies dynamic attention fusion. The performance achieved was roughly 73% for large-scale species. However, the method relies on geo-location data and is not designed for natural images in canonical positions. Hieu et al. [17] employed ViT for embedding, with classification performed using KNN. This simple technique allows for the easy addition of new classes. Unfortunately, as the number of classes or samples increases, it contributes to the time complexity of prediction.

Lee et al. [18] ensembled ViT with ResNet50, DenseNet-201, and Xception networks (convolution-based networks), producing an accuracy of approximately 100%. Even though the performance is flawless, the images in the experimented dataset are scanned images rather than natural ones, which may not be suitable for real-world scenarios in an open-world environment. Dönmez [9] proposed E-ResMLP+ for addressing wheat species classification by employing ResMLP with EfficientNetV2b0. The model produced nearly 99% of the F1-Score. However, this study lacks a comprehensive comparison with available pre-trained models, which needs further investigation. Gustineli et al. [19] employed self-supervised ViT to classify plant species in a multi-label task. This study is still in the preliminary stage of tackling the problem in the PlantCLEF 2024 competition. Nhut et al. [20] employed ViT and BEiT, achieving approximately 99% accuracy for both models, demonstrating the promising performance of ViT.

Unfortunately, limited studies have emerged that enable plant species classification to recognize unknown classes. One of the earliest attempts was conducted by Ghazi et al. [21]. They consolidated GoogLeNet and VGGNet to produce an averaged score to predict known classes. For identifying open set samples, they fine-tuned GoogLeNet as a separate model by solving binary classification problems. Fang et al. [22] combined a CNN for known class prediction with Weighted SVDD for single-class recognition of unknown class prediction. This approach necessitates an extra step to construct the OSR model by separate training through the SVDD model, similar to the study by Ghazi et al. [21], which was conducted about seven years ago. Ment et al. [23] extended ARPL by using ViT pretrained on PlantCLEF2022 and additive margin softmax loss which provided high accuracy and unkown rejection performance. However, the number of classes experimented is low, 6 to 15 which may not encapsulate the performance in open-world where lots of species presents.

Our study experiments with ViT and introduces an alternative to maximum probability or logits using the TKLDS scoring mechanism. Unlike previous approaches that modify the architecture or learning mechanism, our method solely utilizes the logits of trained models. Therefore, TKLDS has the potential to be applied to any existing OSR model.

3. Vision Transformer

The Vision Transformer (ViT) is inspired by the transformer architecture originally developed for machine translation tasks [24]. Unlike traditional convolutional neural networks (CNNs), the ViT model is specifically designed for image classification and omits the decoder component typically used in transformers for sequence generation. A key innovation of ViT is its departure from convolutional operations, instead utilizing Multi-Layer Perceptrons (MLPs) for local feature learning and Multi-Head Self-Attention (MSA) mechanisms for capturing global dependencies. This architecture enables the model to effectively learn and classify images by leveraging the power of self-attention across the entire patches.

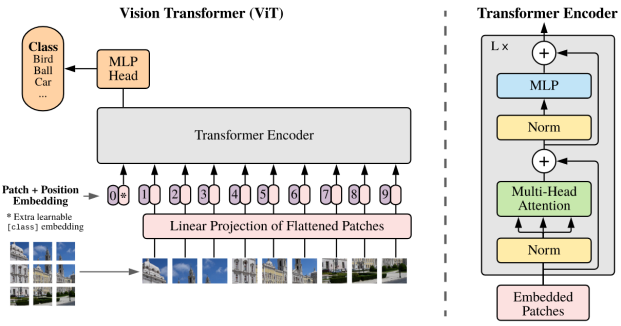


Figure 1: Vision Transformer model [25]

The ViT model, as illustrated by [25], is depicted in Figure 1. In this model, the original image is divided into learnable patches, which are then flattened into feature vectors. These flattened embeddings, before being fed into the transformer encoder, are augmented with learnable positional embeddings as denoted in Equation 1. Here, the notation [...] indicates the concatenation of vectors. The token x_{class}, representing the class embedding, serves as a representation of the image y.

$$z_0 = [x_{class}; x_p^1 E; ...; x_p^N E] + E_{pos},$$
$$E \in \mathbb{R}^{(P^2.C)\times D}, E_{pos} \in \mathbb{R}^{(N+1)\times D} \quad (1)$$

The combination of position embeddings and patch linear projections forms vectors in $\mathbb{R}^{(N+1)\times D}$, which are then forwarded to the Transformer Encoder, as shown in Figure 1. The next latent feature is computed based on the previous layer using Equation 2. Here, the function LN(.) denotes

Layer Normalization, while MSA(.) represents Multi-Head Self-Attention, as defined in Equation 8. Subsequently, the next layer processes $\tilde{z}_l$ through Layer Normalization (LN) followed by a Multi-Layer Perceptron (MLP) with a residual connection, as shown in Equation 3. Finally, the output prediction is obtained by taking the final output of the first patch, which is associated with the class embedding x_{class}, as presented in Equation 4.

$$\tilde{z}_l = MSA(LN(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = MLP(LN(\tilde{z}_l)) + \tilde{z}_l \tag{3}$$

$$y = LN(z_l^0) \tag{4}$$

Back to the Multi-Head Self-Attention (MSA), the process involves learning the Query (q), Key (k), and Value (v) vectors as shown in Equation 5. This calculation can be accelerated through multiplication by U_{qkv}. The importance of the features, represented as attention weights A, is derived from the softmax of the dot product of the query q and key k, as shown in Equation 6. The weights A are then utilized to compute the self-attention through the value v using Equation 7. This process mimics obtaining information in information retrieval, where the query is the input for searching, the key acts as identifiers or features, and the value contains the information.

MSA is essentially the repetition of the self-attention ($SA(.)$) process with k executions, and the results are combined and multiplied by the learnable matrix U_{msa} to maintain the same dimensionality for the next layers, as shown in Equation 8.

$$[q, k, v] = zU_{qkv}, U_{qkv} \in \mathbb{R}^{D \times 3D_h} \tag{5}$$

$$A = softmax(\frac{qk^T}{\sqrt{D_h}}), A \in \mathbb{R}^{N \times N} \tag{6}$$

$$SA(z) = Av \tag{7}$$

$$MSA(z) = [SA_1(z); SA_2(z); ...; SA_k(z)]U_{msa}, \quad U_{msa} \in \mathbb{R}^{k.D_h \times D} \tag{8}$$

To enhance the ability of the Vision Transformer (ViT) to identify unknown classes, we employed transfer learning, using ViT as a feature extractor to train the classifier. Subsequently, the score derived from our proposed function, detailed in the next section, was utilized. The AUROC metric was then employed to assess the model's capability in distinguishing between known and unknown classes, independent of any threshold.

4. Proposed Scoring Function: Top-K Logit Disparity Score (TKLDS)

Many OSR techniques utilize either maximum probability or logit scores exclusively. This approach may be suboptimal, as it does not incorporate additional informative values. To address this limitation, we propose a novel scoring mechanism termed Top-K Logit Disparity Score (TKLDS).

The intuition behind TKLDS is to leverage the disparity among logits rather than relying solely on the maximum logit, thereby enhancing the rejection of unknown classes. We hypothesize that known classes will generate strong activation on one logit while yielding low activation on others, resulting in high disparity. Conversely, unknown classes are likely to activate multiple logits, leading to low disparity. The maximum logit score serves as the basis for calculating the disparity with other logits.

The TKLDS is defined in Equation 9. The vector z represents the sorted logits in descending order, i.e., $z_1 \geq z_2 \geq \ldots \geq z_n$, where n is the number of logits. The parameter k denotes the number of logits used in computing the score. When $k = 1$, only the maximum logit is considered, resulting in no disparity score. For $k > 1$, the mean disparity from z_1 is calculated, indicating the utilization of multiple logits. We will demonstrate in the Results and Discussion section that TKLDS is more effective than using maximum probability or logit alone. Furthermore, TKLDS can be applied to any deep neural network model.

$$TKLDS_k(z) = \begin{cases} z_1 & \text{if } k = 1 \\ \frac{1}{k-1} \sum_{i=2}^{k} (z_1 - z_i) & \text{if } k > 1 \end{cases} \tag{9}$$

5. Experimental Settings

For all experimented models, we utilized the pre-trained models and make them as feature extractor. We set the epoch to 50 with learning rate of 0.001 and the ADAM optimizer. For comparison, we took an epoch with the highest score to unleash each model potential. We constructed each dataset to have five different combinations and reported the mean and standard deviation for each model.

We employ the VNPLANT200 dataset, which represents medicinal plants in Vietnam. VNPLANT200 consists of 20,000 images across 200 different species, with each species contributing 100 images, resulting in a balanced dataset. The images were captured in natural settings.

For the dataset separation, we constructed the set for known classes and unknown classes without overlapping classes between them. We divided the dataset into 5 different combinations in which the first combination use the same order of the classes provided by the dataset. We provided more details at our GitHub repository: https://github.com/gusti-alfarisy/TKLDS

In the first experimentation, we evaluated first the performance on maximum probability and logit for each pre-trained models. Afterwards, we experimented with the proposed TKLDS by tuning the k number from 1 to 5. We observed the performance through CSAAUROC as shown in Equation 10 which is the arithmetic mean between Closed-Set Accuracy (CSA) for known classes and Area Under the Receiver Operating Characteristic curve (AUROC) between known and unknown classes.

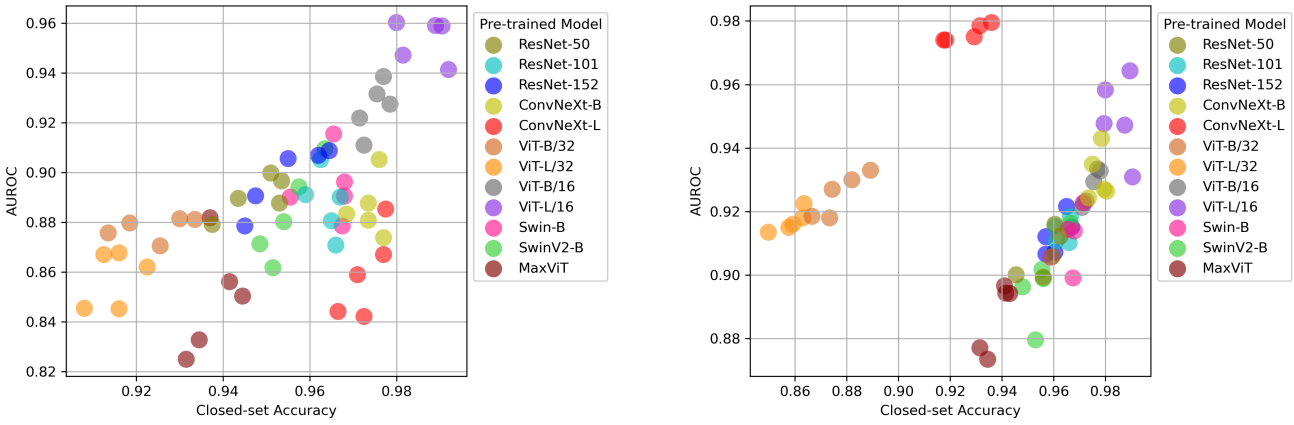
$$CSAAUROC = \frac{(CSA + AUROC)}{2} \tag{10}$$

6. Results and Discussions

Before applying the Top-K Logit Disparity Score (TKLDS) to the output scores, we first evaluated all models using

Table 1: The performance using maximum Logit (L) and maximum Probability (P) accross various pre-trained models

Pre-trained Model	CSA (L)	CSA (P)	AUROC (L)	AUROC (P)	CSAAUROC (L)	CSAAUROC (P)	Best Score
ResNet-50 [26]	0.948 ± 0.007	0.957 ± 0.007	0.891 ± 0.008	0.907 ± 0.007	0.919 ± 0.007	**0.932 ± 0.007**	P
ResNet-101 [26]	0.964 ± 0.003	0.965 ± 0.005	0.888 ± 0.013	0.915 ± 0.005	0.926 ± 0.006	**0.940 ± 0.004**	P
ResNet-152 [26]	0.955 ± 0.009	0.961 ± 0.005	0.898 ± 0.013	0.913 ± 0.006	0.926 ± 0.011	**0.937 ± 0.005**	P
ConvNeXt-B [27]	0.974 ± 0.003	0.977 ± 0.003	0.886 ± 0.012	0.931 ± 0.008	0.930 ± 0.006	**0.954 ± 0.004**	P
ConvNeXt-L [27]	0.973 ± 0.005	0.976 ± 0.003	0.859 ± 0.018	0.927 ± 0.008	0.916 ± 0.011	**0.951 ± 0.005**	P
ViT-B/32 [25]	0.924 ± 0.008	0.925 ± 0.007	0.878 ± 0.005	0.877 ± 0.009	**0.901 ± 0.005**	**0.901 ± 0.008**	L and P
ViT-L/32 [25]	0.915 ± 0.005	0.917 ± 0.003	0.857 ± 0.011	0.859 ± 0.006	0.886 ± 0.007	**0.888 ± 0.004**	P
ViT-B/16 [25]	0.975 ± 0.003	0.975 ± 0.003	0.926 ± 0.010	0.928 ± 0.006	**0.951 ± 0.006**	**0.951 ± 0.004**	L and P
ViT-L/16 [25]	0.987 ± 0.005	0.985 ± 0.005	0.953 ± 0.009	0.950 ± 0.013	**0.970 ± 0.005**	0.968 ± 0.006	L
Swin-B [28]	0.965 ± 0.005	0.966 ± 0.005	0.894 ± 0.014	0.911 ± 0.009	0.930 ± 0.007	**0.939 ± 0.006**	P
SwinV2-B [29]	0.955 ± 0.006	0.956 ± 0.007	0.883 ± 0.019	0.899 ± 0.013	0.919 ± 0.012	**0.927 ± 0.009**	P
MaxViT [30]	0.938 ± 0.005	0.938 ± 0.005	0.849 ± 0.022	0.887 ± 0.011	0.893 ± 0.013	**0.913 ± 0.008**	P



(a) Maximum Logit **(b)** Maximum Probability

Figure 2: Scatter plot depicting the accuracy and AUROC performance metrics for all experiments, with each pre-trained model evaluated using 5 distinct combinations.

maximum logit and probability. This evaluation establishes a baseline performance between the two scores. Through this experimentation, we can observe the impact on logits and probabilities, noting that maximum probability is typically used to reject unknown classes. However, in subsequent experiments, this is not the case with TKLDS. Different pre-trained models were evaluated to assess the consistency of each scoring method across various architectures.

We denote the accuracy of known classes as Closed-Set Accuracy (CSA) to distinguish it from the performance in rejecting unknown classes. Additionally, we define CSAAUROC as the arithmetic mean of the CSA and AUROC scores. For the sake of reproducibility, we published the source code that is available at: https://github.com/gusti-alfarisy/TKLDS.

The performance of the pre-trained models in terms of CSA and AUROC is presented in Table 1. Text in bold style represents the highest CSAAUROC score achieved from either logit or probability, while text in underline style denotes the top five CSAAUROC scores among all pre-trained models.

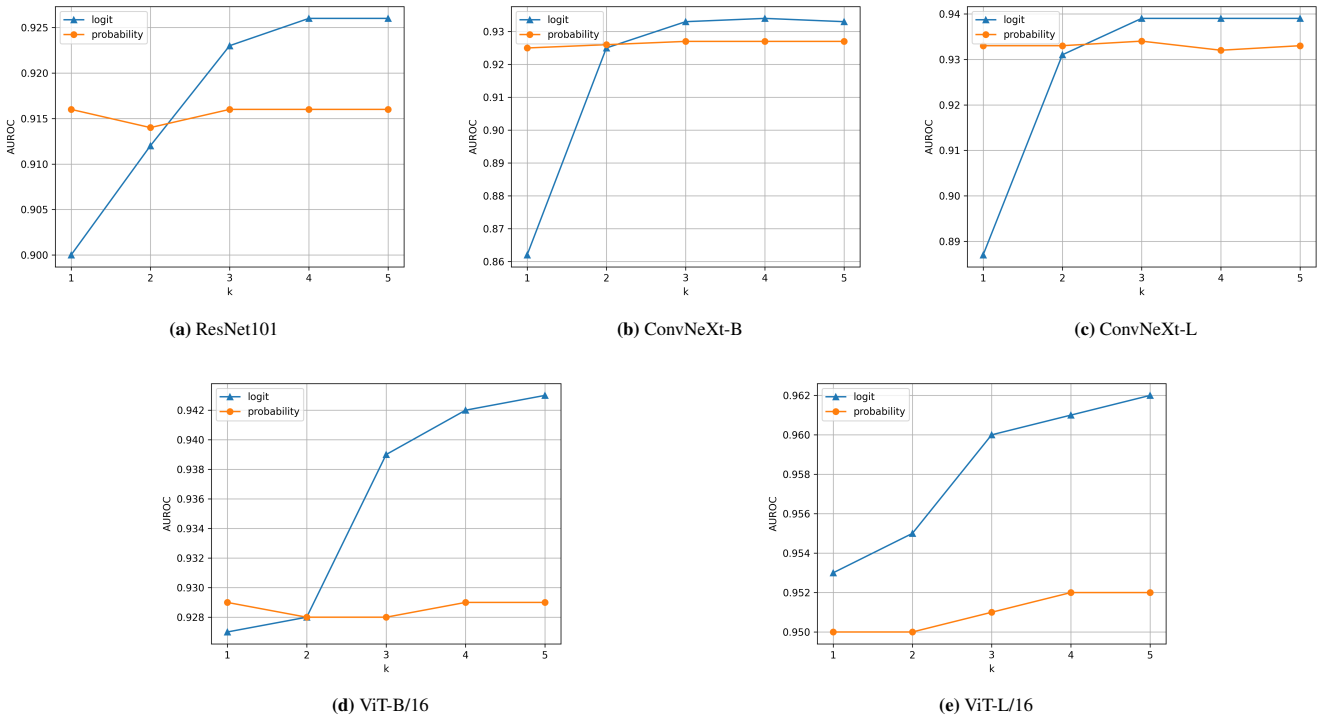
We observe that all pre-trained models achieved a CSA above 90% with both maximum logit and maximum probability, indicating strong performance in known-class prediction. In contrast, the AUROC scores varied, ranging from 0.85 to 0.95. Overall, probability-based evaluation (P) yielded superior performance compared to logit (L) for most pre-trained models.

The capability of open-set recognition models is reflected in both CSA and AUROC; thus, CSAAUROC is a crucial metric for evaluation. From Table 1, it is evident that most pre-trained models achieved a CSAAUROC above 0.9, indicating effective open-set recognition based on the VNPLANT200 dataset. The ViT-L/16 model demonstrated the highest performance with a CSAAUROC score of 0.97 using logit values, with CSA approximately 99% and AUROC around 0.95.

Based on the highest CSAAUROC scores, we selected the top five models for further experimentation with TKLDS. Using these top models, we aim to assess the effectiveness of TKLDS, which is expected to provide reliable performance. This approach also simplifies the comparison process.

Further analysis involves mapping the performance of each combination to its corresponding pre-trained model, as illustrated in Figure 2. Figure 2a depicts the prediction capability using maximum logit, while Figure 2b shows the results using maximum probability.

From Figure 2a, it is evident that ViT-L/16 achieved the highest CSA and AUROC compared to other models. Interestingly, transformer-based networks such as ViT-L/32 and ViT-B/32 exhibited the lowest CSA and AUROC scores. The "32" in ViT indicates that these models use 32x32 pixel patches of the input image, whereas the ViT-L/16 model uses smaller 16x16 pixel patches. Another transformer-based network, MaxViT, demonstrated slightly higher CSA but yielded an extremely low AUROC score in both experimental conditions. In the case of maximum logit, convolution-based networks showed competitive performance relative to transformer-based networks, with the exception of ViT-L/16



(a) ResNet101 **(b)** ConvNeXt-B **(c)** ConvNeXt-L

(d) ViT-B/16 **(e)** ViT-L/16

Figure 3: Top-K Logit Disparity Score (TKLDS) with different k values compared to the disparity of probability using Resnet101, ConvNeXt-B, ConvNeXt-L, ViT-B/16, and ViT-L/16.

and ViT-B/16.

In Figure 2b, employing maximum probability reveals different performance behaviors compared to maximum logit. ConvNeXt-L achieved the highest AUROC score but exhibited lower performance in CSA compared to most other models. Conversely, ViT-L/16 attained the highest CSA but had a lower AUROC score compared to ConvNeXt-L. The models with the largest pixel size, ViT-B/32 and ViT-L/32, demonstrated the worst performance, similar to the results observed with maximum logit.

Interestingly, the base model (ViT-B/32) outperformed the larger model (ViT-L/32), despite having fewer parameters and lower complexity. Additionally, MaxViT did not match the performance of convolution-based models such as ResNet and ConvNeXt. From this figure, it is evident that maximum probability generally indicates higher performance for unknown class rejection, as observed through the AUROC score. However, we will demonstrate that this trend does not hold when TKLDS is applied.

6.1. Top-K Logit Disparity Score versus Probability

The effect of k in the on unknown class rejection performance is illustrated in Figure 3. TKLDS with $k > 2$ consistently outperforms both maximum probability and the same calculation with logit. When using a single output, the logit-based approach generally yields lower performance than the probability-based method (except for ViT-L/16), with a particularly pronounced difference in convolution-based networks. This observation may lead to the biased conclusion that maximum probability or single score is a reliable indicator for the belief system of unknown classes, a perspective commonly adopted in open-set recognition models [31, 32, 15]. Another study utilized the second maximum score for rejecting unknown classes without accounting for other logits [33].

Applying TKLDS with higher k demonstrates an improvement by leveraging top-k logit values as a measure for rejecting unknown classes. This indicates that top-k probability values do not enhance the capability to reject unknown classes. Instead, it suggests that logits provide more valuable information than probabilities. We argue that since probabilities are derived from softmax normalization, important information is lost during this process due to the exponential nature of softmax, which disregards the linear relationships between output values obtained from deep learning models.

Table 2: The best CSAAUROC using TKLDS through maximum Logit (L) and Probability (P)

Pretrained Model	L/P	k	CSAAUROC
ResNet50	L	4	0.937 ± 0.007
ResNet101	L	4	0.945 ± 0.005
ResNet152	L	4	0.942 ± 0.005
ConvNeXt-B	L	5	0.959 ± 0.005
ConvNeXt-L	L	3	0.954 ± 0.006
ViT-B/16	L	5	0.959 ± 0.004
ViT-L/16	**L**	**4**	**0.975 ± 0.005**
ViT-B/32	L	5	0.911 ± 0.008
ViT-L/32	L	4	0.897 ± 0.005
Swin-B	L	4	0.945 ± 0.005
SwinV2-B	L	4	0.935 ± 0.010
MaxViT	L	5	0.921 ± 0.008

The best CSAAUROC scores for all pre-trained models are presented in Table 2. The results clearly indicate that logits are a better indicator than probabilities, as all models achieved the highest CSAAUROC scores using logit values. Furthermore, a higher number of k consistently yielded improved performance, suggesting that even less dominant logits contain valuable information for rejecting unknown

classes. A value of $k \geq 4$ appears to be a good starting point, as it provided the best scores for most pre-trained models, as shown in Table 1.

Regarding pre-trained model performance, ViT-L/16 achieved the highest CSAAUROC score with stable results (with a standard deviation of 0.005). All pre-trained models using TKLDS attained approximately 90% in CSAAUROC, demonstrating robust performance in both known-class classification and unknown-class rejection. In contrast, models with larger patch resolutions, such as ViT-L/32, exhibited the worst performance, whereas the lighter model ViT-B/32 provided competitive results, as detailed in Table 2.

Using two variants of deep learning models (convolutional and transformer-based), the performance was competitive on the VNPLANT200 dataset. Convolution-based networks generally produced higher CSAAUROC scores compared to most transformer-based networks, indicating that many transformer architectures, despite their advanced design, face challenges in open-set recognition tasks with plant species datasets. Nevertheless, the Vision Transformer model ViT-L/16 surpassed all convolution-based networks by approximately 2-4% in terms of CSAAUROC score. For future research, both ViT-L/16 and ViT-B/16 should be considered as baseline models for open plant species recognition tasks.

From this analysis, it is evident that the Top-K Logit Disparity Score (TKLDS) is a simple yet promising scoring function that enhances the model's capability to reject unknown classes. The advantage of this scoring mechanism is its applicability across various types of architectures without dependence on a specific model. We advocate for the use of TKLDS in future comparative analyses for open-set recognition, particularly in the context of plant species classification.

7. Conclusions

In this study, we experimented with plant-species recognition capable of rejecting unknown classes using Vision Transformer. Our results suggest that Vision Transformer should be employed as a baseline for open plant-species recognition. Furthermore, we proposed a scoring function named Top-K Logit Disparity Score (TKLDS) as a primary method to identify unknown classes. We demonstrated that TKLDS improves the model's ability to recognize unknown species compared to maximum probability or logit. TKLDS is a simple yet promising scoring mechanism that can be applied to any deep learning model. Our experiments also indicate that a higher number of k with $k \geq 4$ in TKLDS unlocks the potential of the pre-trained model in an open environment. Future research should explore TKLDS with various deep learning architectures and different values of K to assess its effectiveness in relation to model architecture and the number of classes.

Declaration of Conflicting Interests

The authors declare that there are no competing interests that could have influenced the work of our study.

References

[1] A. K. Verma, P. R. Rout, E. Lee, P. Bhunia, J. Bae, R. Y. Surampalli, T. C. Zhang, R. D. Tyagi, P. Lin, and Y. Chen, "Biodiversity and Sustainability," in *Sustainability*, 1st ed., R. Surampalli, T. Zhang, M. K. Goyal, S. Brar, and R. Tyagi, Eds. Wiley, May 2020, pp. 255–275, doi: 10.1002/9781119434016.ch12.

[2] SIDA, "Urban Development: Biodiversity and Ecosystems," *Swedish International Development Cooperation Agency (SIDA)*, 2016. [Online]. Available: https://cdn.sida.se/publications/files/sida62003en-urban-development-biodiversity-and-ecosystems.pdf

[3] A. Opoku, "Biodiversity and the built environment: Implications for the sustainable development goals (sdgs)," *Resources, Conservation and Recycling*, vol. 141, pp. 1–7, 2019, doi: 10.1016/j.resconrec.2018.10.011.

[4] M. R. Marselle, S. J. Lindley, P. A. Cook, and A. Bonn, "Biodiversity and Health in the Urban Environment," *Current Environmental Health Reports*, vol. 8, no. 2, pp. 146–156, 2021, doi: 10.1007/s40572-021-00313-9.

[5] G. A. W. Rook, "Regulation of the immune system by biodiversity from the natural environment: An ecosystem service essential to health," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 18 360 – 18 367, 2013, doi: 10.1073/pnas.1313731110 .

[6] J. Methorst, A. Bonn, M. Marselle, K. Böhning-Gaese, and K. Rehdanz, "Species richness is positively related to mental health – a study for germany," *Landscape and Urban Planning*, vol. 211, p. 104084, 2021, doi: 10.1016/j.landurbplan.2021.104084.

[7] J. Wäldchen and P. Mäder, "Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review," *Archives of Computational Methods in Engineering*, vol. 25, no. 2, pp. 507–543, Apr. 2018, doi: 10.1007/s11831-016-9206-z.

[8] K. J. Gaston and M. A. O'Neill, "Automated species identification: why not?" *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 359 1444, pp. 655–67, 2004, doi: 10.1098/rstb.2003.1442.

[9] E. Dönmez, "Hybrid convolutional neural network and multilayer perceptron vision transformer model for wheat species classification task: E-ResMLP+," *European Food Research and Technology*, vol. 250, no. 5, pp. 1379–1388, May 2024, doi: 10.1007/s00217-024-04469-0.

[10] A. P. Sundara Sobitha Raj and S. K. Vajravelu, "DDLA: dual deep learning architecture for classification of plant species," *IET Image Processing*, vol. 13, no. 12, pp. 2176–2182, Oct. 2019, doi: 10.1049/iet-ipr.2019.0346.

[11] A. Kaya, A. S. Keceli, C. Catal, H. Y. Yalic, H. Temucin, and B. Tekinerdogan, "Analysis of transfer learning for deep neural network based plant classification models," *Computers and Electronics in Agriculture*, vol. 158, pp. 20–29, Mar. 2019, doi: 10.1016/j.compag.2019.01.041.

[12] M. J. M. Christenhusz and J. W. Byng, "The number of known plants species in the world and its annual increase," *Phytotaxa*, vol. 261, no. 3, pp. 201–217, May 2016, doi: 10.11646/phytotaxa.261.3.1.

[13] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian, "Learning open set network with discriminative reciprocal points," in *ECCV*, 2020, doi: 10.1007/978-3-030-58580-8_30.

[14] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust classification with convolutional prototype learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3474–3482, 2018, doi: 10.1109/CVPR.2018.00366.

[15] G. A. F. Alfarisy, O. A. Malik, and O. W. Hong, "Quad-Channel Contrastive Prototype Networks for Open-Set Recognition in Domain-Specific Tasks," *IEEE Access*, vol. 11, pp. 48 578–48 592, 2023, doi: 10.1109/ACCESS.2023.3275743.

[16] H. Pan, L. Xie, and Z. Wang, "Plant and animal species recognition based on dynamic vision transformer architecture," *Remote Sensing*, vol. 14, no. 20, 2022, doi: 10.3390/rs14205242.

[17] N. V. Hieu, N. L. H. Hien, L. V. Huy, N. H. Tuong, and P. T. K. Thoa, "Plantkvit: A combination model of vision transformer and knn for forest plants classification," *JUCS - Journal of Universal Computer Science*, vol. 29, no. 9, pp. 1069–1089, 2023, doi: 10.3897/jucs.94657.

[18] C. P. Lee, K. M. Lim, Y. X. Song, and A. Alqahtani, "Plant-cnn-vit: Plant classification with ensemble of convolutional neural networks and vision transformer," *Plants*, vol. 12, no. 14, 2023, doi: 10.3390/plants12142642.

[19] M. Gustineli, A. Miyaguchi, and I. Stalter, "Multi-label plant species classification with self-supervised vision transformers," 2024. [Online]. Available: https://arxiv.org/abs/2407.06298

[20] D. T. N. Nhut, T. D. Tan, T. N. Quoc, and V. T. Hoang, "Medicinal plant recognition based on vision transformer and beit," *Procedia Computer Science*, vol. 234, pp. 188–195, 2024, doi: 10.1016/j.procs.2024.02.165.

[21] *Open-set Plant Identification Using an Ensemble of Deep Convolutional Neural Networks*, 2016. [Online]. Available: https://research.sabanciuniv.edu/id/eprint/29408/

[22] T. Fang, Z. Li, J. Zhang, D. Qi, and L. Zhang, "Open-Set Recognition of Wood Species Based on Deep Learning Feature Extraction Using Leaves," *Journal of Imaging*, vol. 9, no. 8, p. 154, Aug. 2023, doi: 10.3390/jimaging9080154.

[23] Y. Meng, M. Xu, H. Kim, S. Yoon, Y. Jeong, and D. S. Park, "Known and unknown class recognition on plant species and diseases," *Computers and Electronics in Agriculture*, vol. 215, p. 108408, Dec. 2023, doi: 10.1016/j.compag.2023.108408.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, doi: 10.1109/CVPR.2016.90.

[27] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 976–11 986, doi: 10.1109/CVPR52688.2022.01167.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022, doi: 10.1109/ICCV48922.2021.00986.

[29] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 12 009–12 019, doi: 10.1109/CVPR52688.2022.01170.

[30] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. C. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *European Conference on Computer Vision*, 2022, doi: 10.1007/978-3-031-20053-3_27.

[31] J. Jang and C. O. Kim, "Teacher-Explorer-Student Learning: A Novel Learning Method for Open Set Recognition," Mar. 2021, doi: 10.1109/TNNLS.2023.3336799.

[32] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust Classification with Convolutional Prototype Learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3474–3482, doi: 10.1109/CVPR.2018.00366.

[33] Y. Shu, Y. Shi, Y. Wang, T. Huang, and Y. Tian, "P-ODN: Prototype-based Open Deep Network for Open Set Recognition," *Scientific Reports*, vol. 10, no. 1, p. 7146, Dec. 2020, doi: 10.1038/s41598-020-63649-6.