MALAYSIAN JOURNAL OF SCIENCE
MJS
ISSN 1394-3065

Regular Issue

Malaysian Journal of Science 44 (4): 18-30 (December 2025) https://mjs.um.edu.my

Enhancing Multispectral Land Use and Land Cover Classification with Transfer Learning and 3D ResNet

Farah Adila Ahmad Marzuki[1a], Helmi Zulhaidi Mohd Shafri[2a*], Siti Nur Aliaa Roslan[3a], Yuhao Ang[4a], Mohammed Mustafa Al-Habshi[5a], Yang Ping Lee[6b], Shahrul Azman Bakar[7b], Haryati Abidin[8b], Hwee San Lim[9c] and Rosni Abdullah[10d]

Abstract: Recent advances in land use and land cover (LULC) classification with remote sensing imagery are driven by state-of-the-art models such as Convolutional Neural Networks (CNNs). Advanced CNN architecture like ResNet can enhance overall classification performance by incorporating residual skip connections. The integration of 3D feature extraction and ResNet architecture suggests a potential improvement in classification tasks. This paper explores the potential of the 3D ResNet model for LULC classification, comparing it with baseline approaches (Support Vector Machine, Random Forest, XGBoost, 1D CNN, 3D CNN) and state-of-the-art 3D models (3D VGG, 3D DenseNet) using WorldView-2 satellite imagery. The 3D ResNet-18 model, fine-tuned via transfer learning on multispectral images, demonstrates significant improvements in classification performance over machine learning models. It achieves the highest Overall Accuracy (OA) of 99.66% and Kappa Accuracy (KA) of 99.39% on the primary dataset. Despite having slightly lower performance on the external validation dataset (OA:82.89%, KA:80.05%) than 3D DenseNet, it is highly efficient with processing times of 490.2 minutes and 3.6 minutes for both datasets respectively. McNemar's test results show 3D ResNet and 3D DenseNet have significant differences in classification performance ($p<0.05$) against other models consistently for both datasets.

Keywords: *Convolutional neural network, deep learning, LULC, multispectral, transfer learning.*

1. Introduction

Land use and land cover (LULC) classification serves an important function as it can be used to identify different types of natural and economic operations which are later used to facilitate better decision-making (Wang et al., 2022). LULC data extracted from remote sensing images are commonly used nowadays due to increasing number of multispectral and hyperspectral satellites along with publicly available dataset. They can be very useful for urban planning and agricultural management (Shaharum et al., 2020). The distinct spectral characteristics of various land cover types across multiple spectral bands are effectively captured by multispectral satellite data or imagery, making it a common choice for LULC classification, which enables more precise classification.

Conventional machine learning (ML) models like Support Vector Machine (SVM), Random Forest (RF) and Extreme Gradient Boost (XGBoost) have been employed extensively throughout the years since they have lower computational complexity and higher interpretability (Sheykhmousa et al., 2020). These models can produce reliable results without much hyperparameter tuning. Contradicting results were reported by several research when comparing these models for LULC classification. Previous studies showed that SVM was able to get a higher accuracy than RF and XGBoost by using Sentinel-2 (Abdi, 2020) satellite imagery. However, findings from another research indicated that RF outperformed SVM for large area mapping using a WorldView-2 image (Jombo et al., 2020). Another research has also shown that XGBoost achieved higher accuracy over SVM and RF for both aerial image and WorldView-2 image (Jozdani et al., 2019).

Deep learning (DL) approaches, particularly Convolutional Neural Networks (CNN), have gained attention for their ability to capture complex patterns in remote sensing images. Deeper CNN networks can learn more intricate details, but this does not always lead to higher accuracy due to the vanishing gradient problem which can hinder training convergence (Noh, 2021). Advanced CNN architectures like Residual Network (ResNet) are later introduced to address this limitation of regular CNN models (He et al., 2015). ResNet has shown exceptional performance in LULC classification using multispectral images (Tong et al., 2020). Complex architectures like ResNet are often used with transfer learning, allowing knowledge from one task to be applied to another. This helps improve model generalization and reduces training time.

Authors information:
[a]Department of Civil Engineering and Geospatial Information Science Research Centre (GISRC), Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, MALAYSIA. Email: frhdila13@gmail.com[1], helmi@upm.edu.my[2], aliaa_roslan@upm.edu.my[3], vincentangkhun@gmail.com[4], alhabshi3k@gmail.com[5]
[b]Geoinformatics Unit, FGV R&D Sdn Bhd, FGV Innovation Centre, PT23417, Lengkuk Teknologi, 71760 Bandar Enstek, Negeri Sembilan, MALAYSIA. Email: yangp.lee@fgvholdings.com[6], shahrul.b@fgvholdings.com[7], haryati.a@fgvholdings.com[8]
[c]School of Physics, Universiti Sains Malaysia (USM), 11800 Gelugor, Penang, MALAYSIA. Email: hslim@usm.my[9]
[d]School of Computer Sciences, Universiti Sains Malaysia (USM), 11800 Gelugor, Penang, MALAYSIA. Email: rosni@usm.my[10]
*Corresponding Author: helmi@upm.edu.my

Received: July, 2024
Accepted: June, 2025
Published: December, 2025

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

Recent research has shown the effectiveness of 3D ResNet models for various applications like medical imaging informatics (Ebrahimi et al., 2020). By harnessing the benefits of 3D feature extraction and residual skip connections, 3D ResNet models offer promising results for accurate land cover classification. Hence, various studies have employed 3D ResNet models for LULC classification with hyperspectral images (Firat et al., 2023). To the extent of research conducted, it is believed that there is no other existing research on the application of 3D ResNet models for multispectral LULC classification.

In summary, the objectives of this paper are:

- To investigate the potential application of deep residual model, 3D ResNet-18 in LULC classification of WorldView-2 multispectral satellite imagery. The model performance was compared against several baseline models including RF, SVM, XGBoost, 1D CNN, and 3D models (3D CNN, 3D VGG and 3D DenseNet).
- To demonstrate the benefits of transfer learning by fine-tuning a pre-trained 3D ResNet-18 model on multispectral satellite imagery.
- To employ WorldView-2 spectral ground truth dataset that provides additional bands that capture detailed spectral information beyond the visible spectrum.

2. Materials and Methods

Study Area

The study was conducted on two oil palm regions (Figure 1) situated in Jerantut, Pahang. The dataset features a broad range of land cover types and oil palm trees at multiple growth stages, which was suitable for testing model robustness. The image (10790 x 10351 pixels) was obtained using WorldView-2, a multispectral satellite imagery. Due to its spectral range of 450 to 800 nm and spatial resolution of 0.3 m post-pansharpening, the image offers a more detailed perspective on land cover features.

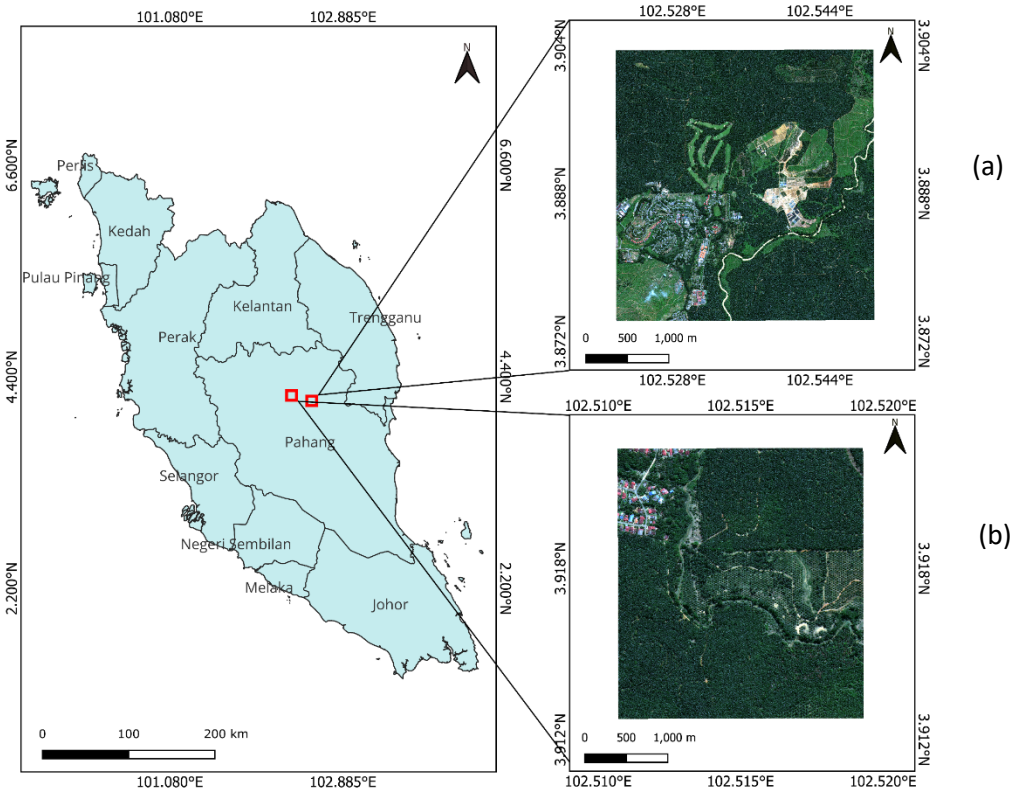

Figure 1. WorldView-2 satellite image of the oil palm plantation located in Jerantut, Pahang, Malaysia, (a) Primary dataset; (b) External validation dataset.

Design of Study

The experiment followed the methodology in Figure 2. Multispectral satellite images were acquired, and seven map classes were defined for classification. A total of 3476175 samples (primary dataset) and 24113 samples (validation dataset) were collected through ground truthing using QGIS software. Stratified k-fold cross-validation (SKCV) was used for data sampling. Three ML models (SVM, RF, XGBoost) and five DL models (1D CNN, 3D CNN, 3D VGG, 3D DenseNet, 3D ResNet) were chosen for LULC classification. Model training and hyperparameter tuning identified the best parameters for each model. Model performance was evaluated using Overall Accuracy (OA), Kappa Accuracy (KA), Precision, Recall, and F1-score.

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

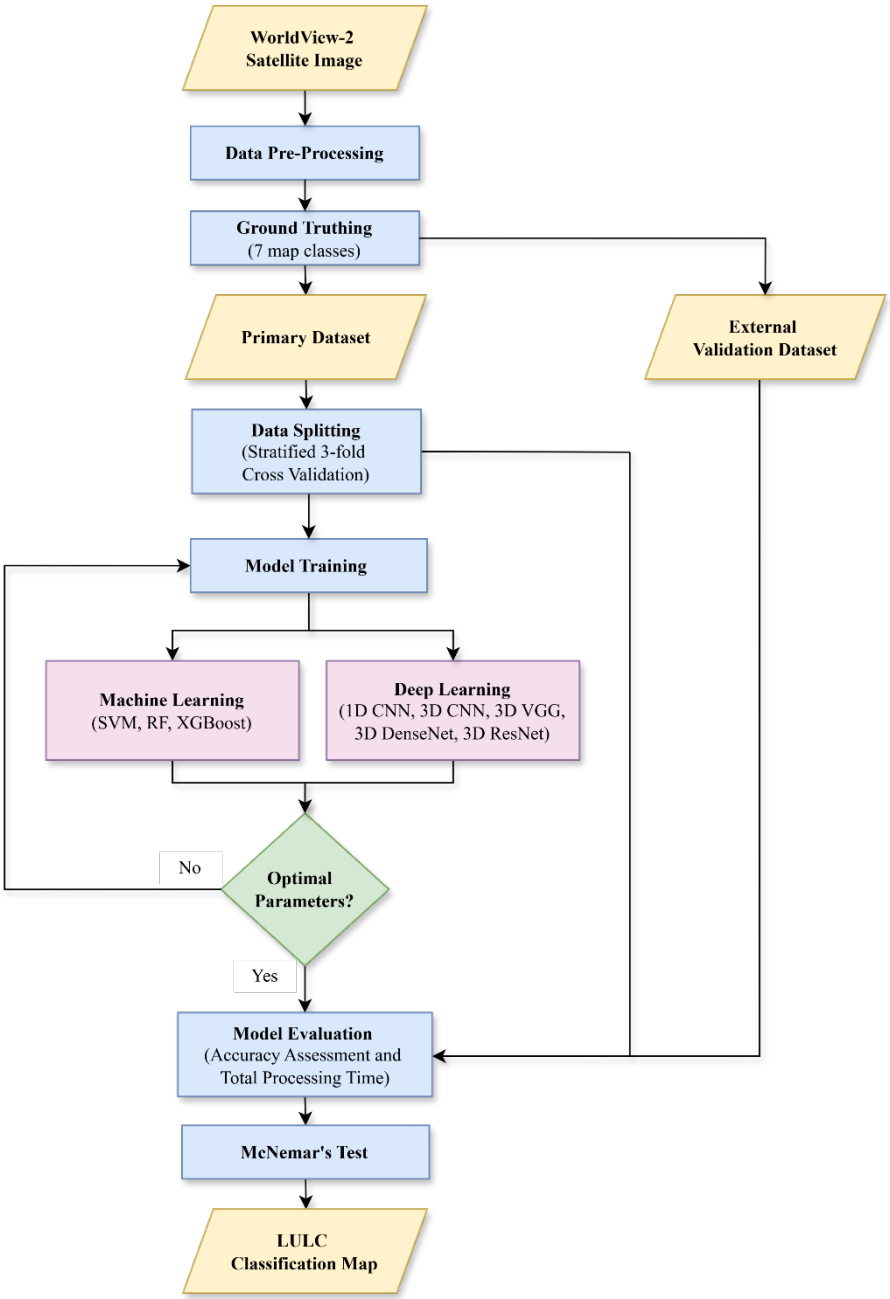

Figure 2. Methodology flowchart.

Sampling Strategy

This study used stratified random sampling and SKCV approach was applied due to the dataset imbalance in Table 1. The SKCV approach is a modified version of normal k-fold cross validation. In the normal k-fold, it splits the dataset randomly into k equal parts. However, the SKCV ensures the same class distribution in every fold like the original dataset; thus, reducing bias during model evaluation. Research performed by Thölke et al. (2023) has shown that SKCV is less sensitive to imbalanced data compared to normal k-fold as it prevents any fold from containing only one class, resulting in a fairer model evaluation.

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

Table 1. Data distribution for primary dataset and external validation dataset.

Class ID	Land Cover Class	Class Description	Number of pixels	
			Primary Dataset	Validation Dataset
1	Water bodies	Natural and artificial water bodies	104052	3734
2	Built-up	Man-made infrastructure	192764	3273
3	Bare soil	Minimal or no vegetation	338023	3748
4	Forest	Open and closed canopy forest	236787	3517
5	Grass	Herbaceous plant cover	237140	3027
6	Oil palm	Oil palm trees of various ages	2222685	3243
7	Other vegetations	Includes other plants like pandan coconut and durian trees	144724	3571
		TOTAL	3476175	24113

Model Architectures

The introduction of ResNet architecture leads to significant improvement in land cover classification as they can utilize skip connections, allowing the training of much deeper models and reducing the optimization difficulty (Tong et al., 2020). By incorporating 3D convolutional filters, 3D ResNet models can be developed to extract spectral-spatial data from remote sensing images. The network depth of 18 layers provides a good balance between the model complexity and computational cost, which makes it a suitable choice.

To reduce overfitting in the 3D ResNet-18 model, we used transfer learning with Dimension Expansion Weight Transfer (DEWT) technique. DEWT is a simple yet effective method, where pre-trained 2D weights are transferred by replicating them across the third dimension. By fine-tuning the model with pre-trained ImageNet weights and adding a new classification layer for the seven classes shown in Figure 3, 3D ResNet-18 model performance and generalizability can be improved. The implementation code is available on GitHub: https://github.com/russelrk/Pre_Trained_3D_CNN. To assess its performance, other state-of-the-art 3D models including 3D VGG-16 (Simonyan & Zisserman, 2015) and 3D DenseNet-121 (G. Huang et al., 2018) were also incorporated in this study. The same type of weight transfer was also used for these models.

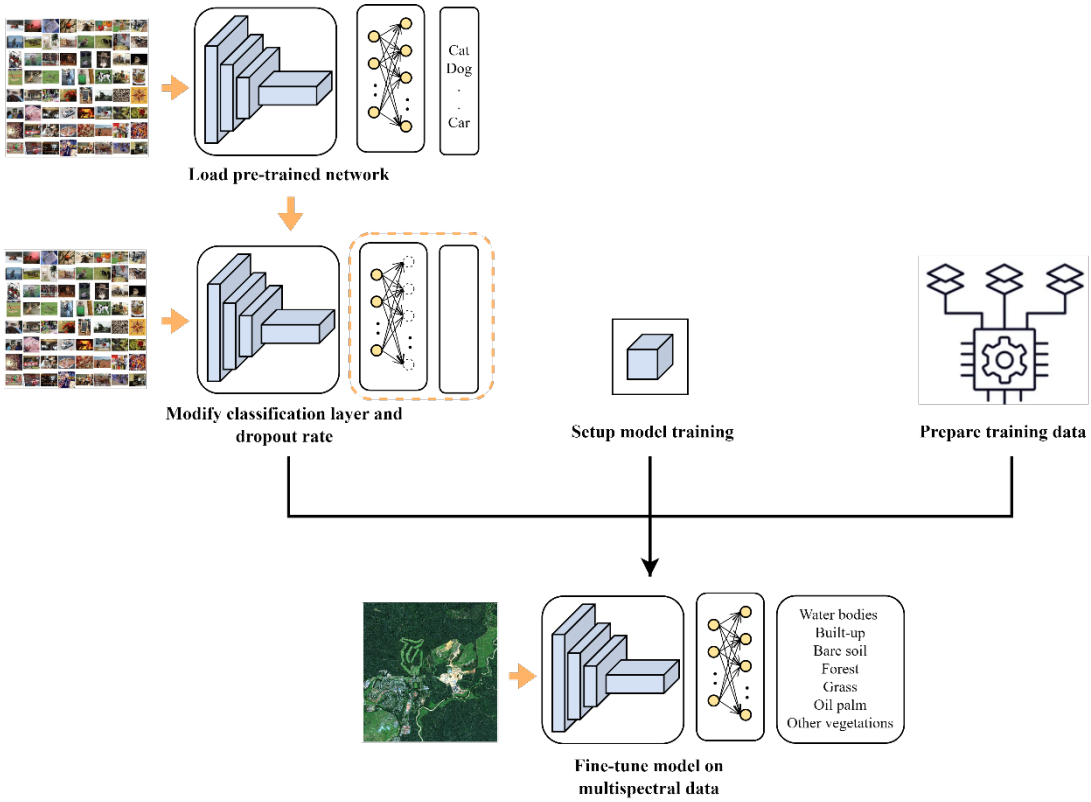

Figure 3. Process of fine-tuning 3D ResNet model.

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

SVM is a popular classifier for LULC classification, especially with small datasets (Vali et al., 2020). It aims to find the best decision boundary that separates target classes by solving a quadratic optimization problem (Sheykhmousa et al., 2020). Due to limited computational resources, linear SVM was used in this study. RF is an ensemble method that improves accuracy by combining multiple decision trees. It performs effectively with high-dimensional data and less prone to overfitting. Like RF, XGBoost uses ensemble learning. It is an improved version of the gradient boosting machine (GBM) algorithm, enhancing both performance and speed (Abdi, 2020). It builds a sequential model with shallow decision trees, optimizing a loss function, while adding regularization to reduce overfitting.

Model Training and Hyperparameter Tuning

All experiments were implemented in Google Colab Pro+ with Python. The processing for ML models was performed on CPU whereas the processing for DL models was performed on A100 GPU NVIDIA with 40 GB of RAM. Hyperparameter tuning was conducted using coarse grid search to determine the best parameters for each model as shown in Table 2. By using PyTorch framework, each DL model was trained in batches of 256 for 100 epochs per fold, with a learning rate of 0.0001. Regularization techniques like using dropout rate of 0.2 and early stopping rounds of 5 were applied to reduce overfitting. The Adam optimizer was used for faster convergence, and a 5x5 pixels patch size was chosen for computational efficiency.

Table 2. Best model parameters after hyperparameter tuning.

Classifier	Parameters
SVM (Linear)	C = 1
RF	n_estimators = 300 max_depth = 30 min_samples_leaf = 1 min_samples_split= 2
XGBoost	n_estimators = 500 max_depth = 7 min_child_weight = 3 subsample = 0.7
1D CNN 3D CNN 3D VGG 3D DenseNet 3D ResNet	learning_rate = 0.0001 batch_size = 256 epochs = 100

Performance Evaluation

Model evaluation was conducted in terms of accuracy and processing times. OA is the proportion of accurately classified instances relative to the overall number of reference instances. KA is measure of agreement between reference data and classifier with the range of -1 to 1. Although these metrics are significant in model evaluation, it can show misleading results in cases of class imbalance; hence, additional metrics were also applied. Precision, as shown in Eq. (1) represents the proportion of correctly classified instances relative to the overall number of instances classified, where TP, FP, and FN represent the number of true positives, false positives, and false negatives respectively.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad (1)$$

Recall in Eq. (2), is the proportion of correctly classified instances to the actual instances.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad (2)$$

F1-score, shown in Eq. (3) is the harmonic mean of Precision and Recall, an indicator of the predictive ability of the model.

$$\text{F1} - \text{score} = 2\left[\frac{(\text{Precision}\,.\,\text{Recall})}{(\text{Precision}+\text{Recall})}\right] \quad (3)$$

McNemar's test was also performed to inspect the statistical significance of difference in the classification performance between models. A 2 x 2 contingency matrix was constructed and calculated to determine the value of chi-square and p-value. The calculation for McNemar's test value is shown in Eq. (4), where f_{12} and f_{21} represent the quantity of accurately classified and misclassified samples respectively. A significance threshold (α=0.05) was fixed before computing the p-value. The null hypothesis (H_0) assumes equal classification performance between models, while the alternative hypothesis (H_1) suggests a difference. H0 is rejected if the p-value is below the threshold.

$$\chi^2 = \frac{(|f_{12}-f_{21}|-1)^2}{(f_{12}+f_{21})} \quad (4)$$

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

3. Results and Discussion

Accuracy Assessment and Processing Time

Based on Figure 4, while the loss curve of the 3D ResNet-18 generally stabilizes, slight fluctuations can be observed in the validation loss. This may result from Adam optimizer's adaptive nature (Kingma & Ba, 2017) and minor imbalances in feature distribution. Experimenting with different optimizers could help address these fluctuations. Despite this, the minimal divergence between training and validation losses indicates no overfitting.

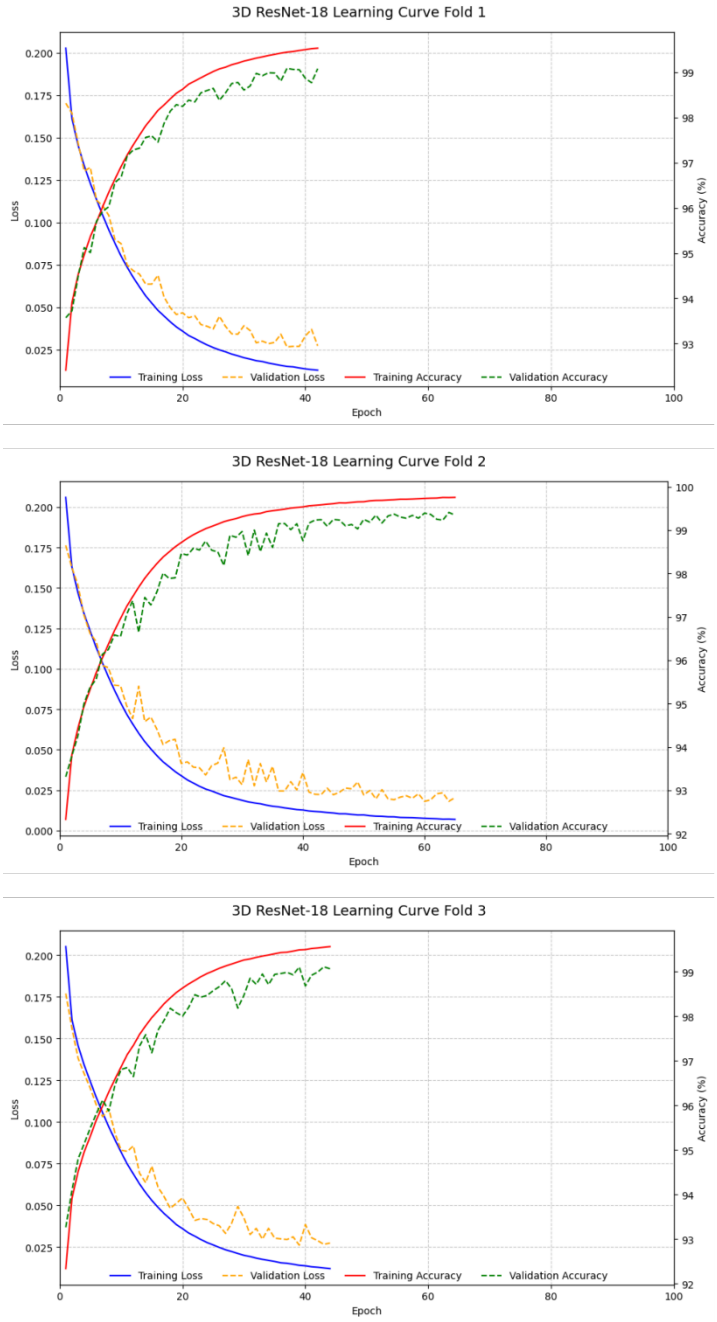

Figure 4. Loss and accuracy learning curves for 3D ResNet.

Overall, 3D ResNet-18 outperformed the other models by achieving a high OA of 99.66% and KA of 99.39% on the primary dataset as shown in Figure 5. Although 3D ResNet-18 has similar ability to 3D CNN in extracting spectral-spatial features, it incorporates residual skip connections which allows it to preserve important information and gradient. The skip connections in ResNet provide a direct path from the input of one layer to the output of deeper layers, allowing for better feature retention in the model. The skip connections also enabled more effective gradient flow as it can bypass certain intermediate layers during backpropagation, addressing the vanishing gradient problem and ensuring more effective weight updates. This is aligned with previous research involving comparison between traditional CNN architecture and ResNet architecture (D & Bhavani, 2023).

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

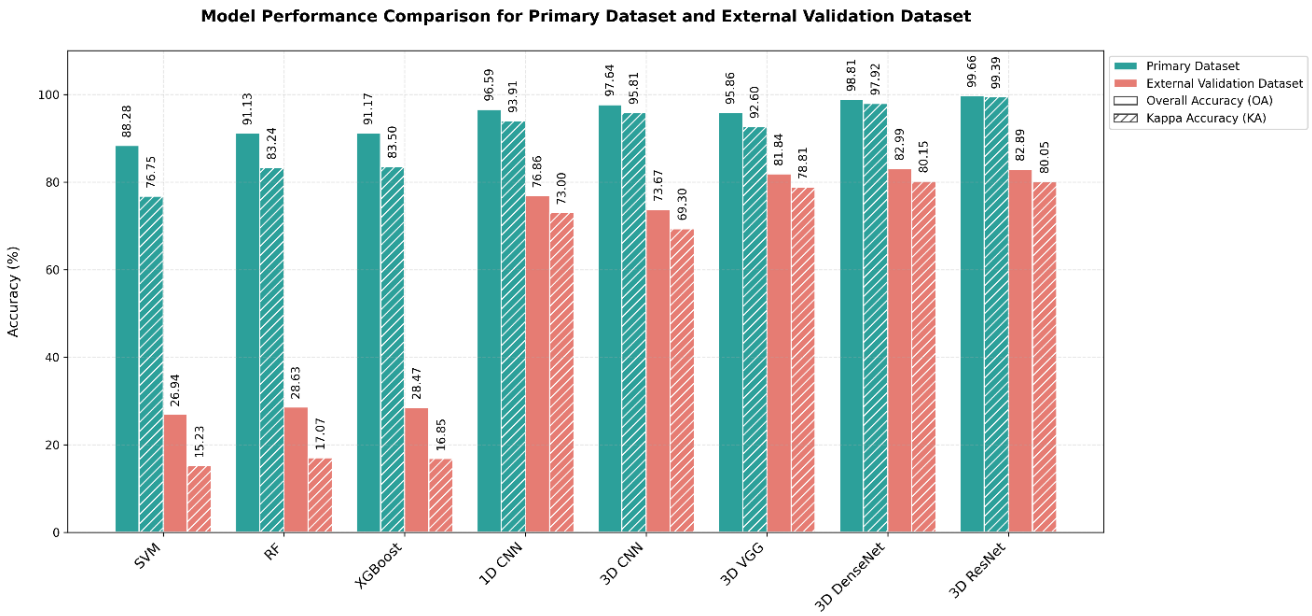

Figure 5. Comparison of model performance (OA and KA) for primary dataset and external validation dataset.

Although 3D ResNet model outperformed both ML models and DL models on the primary dataset, it achieved slightly lower OA (82.89%) and KA (80.05%) than 3D DenseNet for the external validation dataset. Based on the results, the generalizability of the models is clearly impacted when evaluated on the external validation dataset, as seen by the decrease in performance across most models in terms of KA. It also suggests that while DL models like 3D ResNet offer high accuracy, they may require additional strategies such as domain adaptation for better generalization to new areas.

In line with theoretical predictions, 1D CNN has a slightly lower OA and KA than 3D CNN for primary dataset since it only has the ability of capturing the spectral features. Similar results were also reported in other research involving LULC classification using Indian Pines and Wuhan University datasets (Liu et al., 2023). However, 1D CNN has better generalization capability compared to 3D CNN which is probably due to its simpler architecture. There is a significant difference between the classification performance of SVM compared to the other classifiers which is probably due to severe effects of imbalanced datasets to the SVM margin computation.

For primary dataset, all models performed well on most classes with high precision, recall and F1-score (Table 3). However, ML models struggled with forest and other vegetation classes, displaying significant misclassifications. This suggests that ML models may have difficulty differentiating between spectrally similar vegetation types, possibly due to overlapping spectral or spatial characteristics. In contrast, DL models demonstrated improved performance across all classes, especially for challenging vegetation features, indicating their ability to capture more complex feature representations.

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

Table 3. Comparison of model performance (Precision, Recall, F1-score) for primary dataset.

		Classifier							
Metric	**Map Class**	**SVM**	**RF**	**XGBoost**	**1D CNN**	**3D CNN**	**3D VGG**	**3D Dense Net**	**3D ResNet**
Precision	Water bodies	0.997	**1.000**	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**
	Built-up	0.964	0.998	0.998	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Bare soil	0.964	0.999	0.999	**1.000**	**1.000**	0.999	**1.000**	**1.000**
	Forest	0.028	0.760	0.720	0.897	0.916	0.900	0.918	0.986
	Grass	0.962	0.992	0.995	**1.000**	**1.000**	0.998	**1.000**	**1.000**
	Oil palm	0.852	0.893	0.899	0.967	0.977	0.962	0.995	0.997
	Other vegetations	0.000	0.661	0.611	0.839	0.916	0.749	0.933	0.989
Recall	Water bodies	0.944	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Built-up	0.962	0.997	0.997	0.999	**1.000**	0.998	**1.000**	**1.000**
	Bare soil	0.983	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Forest	0.000	0.315	0.342	0.766	0.844	0.703	0.971	0.982
	Grass	0.986	0.994	0.996	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Oil palm	0.996	0.983	0.976	0.985	0.989	0.981	0.987	0.998
	Other vegetations	0.000	0.275	0.330	0.795	0.852	0.785	0.958	0.975
F1-score	Water bodies	0.970	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Built-up	0.963	0.997	0.998	**1.000**	**1.000**	0.999	**1.000**	**1.000**
	Bare soil	0.973	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Forest	0.000	0.445	0.464	0.826	0.879	0.789	0.944	0.984
	Grass	0.974	0.993	0.996	**1.000**	**1.000**	0.999	**1.000**	**1.000**
	Oil palm	0.918	0.936	0.936	0.976	0.983	0.971	0.991	0.997
	Other vegetations	0.000	0.388	0.429	0.817	0.883	0.767	0.945	0.982

Based on the validation results in Table 4, ML model performance dropped significantly across most classes, indicating the struggle with generalization to new geographic areas and environmental conditions. In contrast, DL models achieved strong generalization for all classes with only a slight decrease in accuracy. This highlights the importance for model robustness in different geographical settings.

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

Table 4. Comparison of model performance (Precision, Recall, F1-score) for external validation dataset.

		Classifier							
Metric	**Map Class**	**SVM**	**RF**	**XGBoost**	**1D CNN**	**3D CNN**	**3D VGG**	**3D Dense Net**	**3D ResNet**
Precision	Water bodies	**1.000**	0.000	0.000	0.715	0.733	0.821	**1.000**	**1.000**
	Built-up	**1.000**	0.586	0.583	0.915	0.882	0.953	0.960	0.957
	Bare soil	0.303	0.091	0.139	0.943	0.693	0.950	0.983	0.977
	Forest	0.000	0.698	0.639	0.907	0.920	0.923	0.935	0.970
	Grass	0.062	0.630	0.000	0.998	0.971	0.998	0.985	0.971
	Oil palm	0.192	0.201	0.206	0.512	0.489	0.551	0.487	0.483
	Other vegetations	0.000	0.015	0.011	0.738	0.772	0.767	0.697	0.795
Recall	Water bodies	0.017	0.000	0.000	0.994	0.780	**1.000**	**1.000**	0.999
	Built-up	0.577	0.527	0.493	0.998	**1.000**	**1.000**	**1.000**	**1.000**
	Bare soil	0.336	0.095	0.153	0.541	0.583	0.709	0.915	0.923
	Forest	0.000	0.541	0.515	0.819	0.748	0.773	0.782	0.673
	Grass	0.015	0.010	0.000	0.692	0.727	0.797	0.735	0.738
	Oil palm	**1.000**	0.891	0.883	0.904	0.922	0.926	0.837	0.940
	Other vegetations	0.000	0.001	0.001	0.454	0.441	0.542	0.528	0.526
F1-score	Water bodies	0.033	0.000	0.000	0.831	0.756	0.902	**1.000**	0.999
	Built-up	0.731	0.555	0.534	0.955	0.937	0.976	0.980	0.978
	Bare soil	0.318	0.093	0.146	0.687	0.633	0.812	0.947	0.949
	Forest	0.000	0.609	0.570	0.861	0.825	0.841	0.852	0.794
	Grass	0.024	0.019	0.000	0.817	0.832	0.886	0.842	0.839
	Oil palm	0.322	0.328	0.334	0.653	0.639	0.691	0.615	0.638
	Other vegetations	0.000	0.001	0.003	0.562	0.561	0.636	0.600	0.633

In Table 5, ML models had significantly shorter processing times compared to DL models, with SVM taking only 1.8 minutes for primary dataset and 0.2 minutes for external dataset. In contrast, 3D models require much longer processing times, with 3D DenseNet being the most computationally expensive (1642.8 minutes for primary dataset). These results highlight the trade-off between model complexity and computational efficiency, where more complex 3D models demand more resources. In comparison to 3D VGG and 3D DenseNet, 3D ResNet was the most computationally efficient in LULC mapping for both primary dataset (490.2 minutes) and validation dataset (3.6 minutes).

Table 5. Total processing time for primary dataset and external validation dataset.

	Total Processing Time (min)	
Model	**Primary Dataset**	**External Validation Dataset**
SVM	1.8	0.2
RF	385.2	17.4
XGBoost	41.4	20.2
1D CNN	444.0	1.4
3D CNN	403.2	1.2
3D VGG	1101.6	11.6
3D DenseNet	1642.8	23.3
3D ResNet	490.2	3.6

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

Land Cover Classification Map

Based on Figure 6 and Figure 7, all models are capable of classifying land cover classes like water bodies, built-up regions, bare soil, and grass with high accuracy due to their distinct spectral signatures. Despite both rivers and ponds being categorized under the same class, the ML models could only classify rivers accurately and not ponds. Misclassifications often occurred between ponds and built-up areas, likely due to the similar spectral signatures of herbaceous vegetation in the ponds and features of the built-up areas, as well as their proximity to each other. It appears that 3D ResNet and 3D DenseNet have the least misclassifications between ponds and built-up areas, signifying their greater ability in capturing the subtle variations in spectral and spatial features.

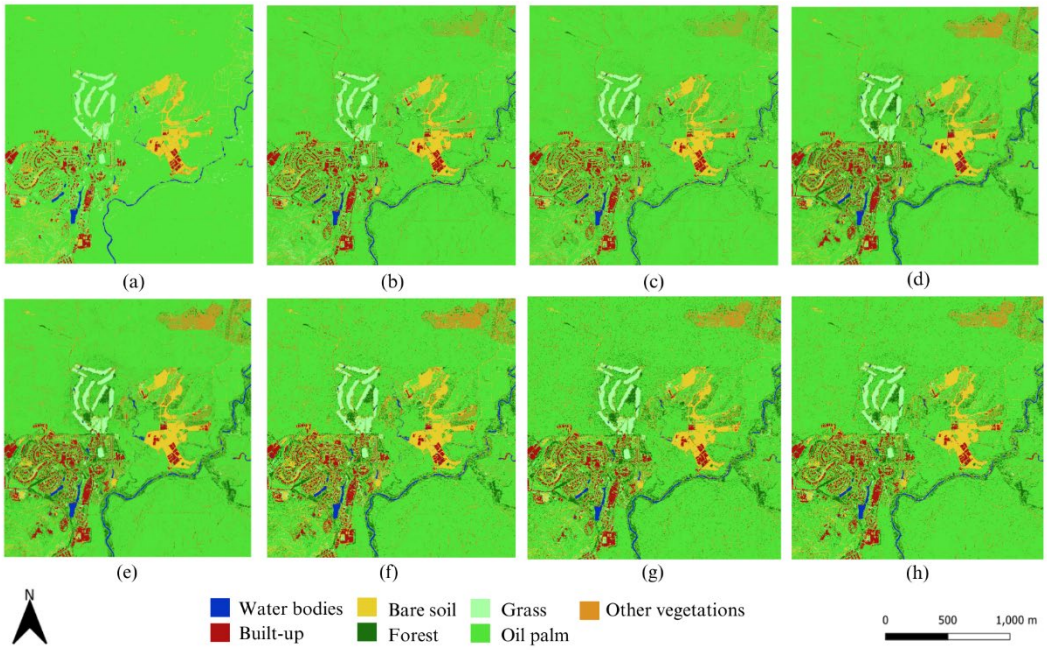

Figure 6. Classification results of primary dataset with (a) SVM; (b) RF; (c) XGBoost; (d) 1D CNN; (e) 3D CNN; (f) 3D VGG; (g) 3D DenseNet; (h) 3D ResNet.

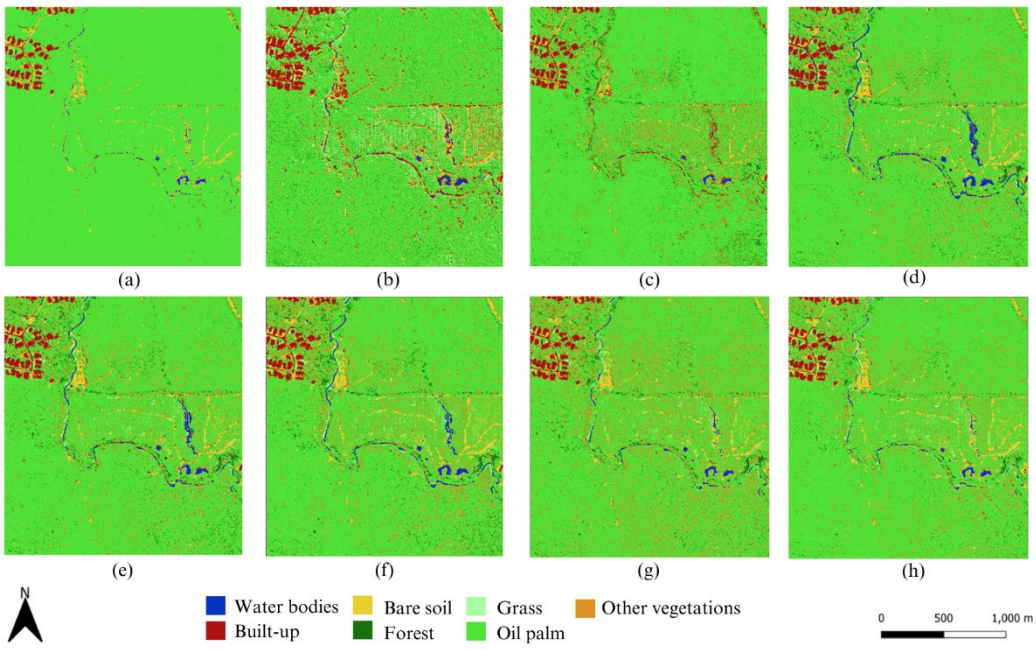

Figure 7. Classification results of external validation dataset with (a) SVM; (b) RF; (c) XGBoost; (d) 1D CNN; (e) 3D CNN; (f) 3D VGG; (g) 3D DenseNet; (h) 3D ResNet.

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

High intra-class variation was observed in built-up, oil palm, and other vegetation samples. Built-up areas varied due to material and structural differences, while oil palm variation resulted from planting density and age, affecting spectral and textural patterns. All models classified built-up and oil palm samples well, but other vegetation samples were more challenging due to mixed patterns. Accurate classification of other vegetation was achieved with 3D ResNet and 3D DenseNet. Despite good intra-class similarity, classifying forest samples was challenging for ML models, which struggled to distinguish forest from oil palm. DL models, especially 3D ResNet and 3D DenseNet performed better as CNNs can extract vegetation features effectively.

McNemar's Test

McNemar's test revealed critical insights into model robustness for primary dataset (Table 6) and external validation dataset (Table 7). 3D ResNet and 3D DenseNet demonstrated consistent superiority (p<0.05) over ML models across datasets, outperforming other DL models. While 1D CNN showed significant advantages over ML models (p<0.05), its generalizability was less robust on external data compared to 3D ResNet and 3D DenseNet. Similarly, both 3D CNN and 3D VGG displayed inconsistencies in model performance across datasets. The variability observed in simpler (1D CNN) or less optimized 3D architectures (3D CNN, 3D VGG) further emphasizes the critical role of architectural design in ensuring generalizability. 3D ResNet evidently achieved comparable classification performance to 3D DenseNet (p = 1.000) with shorter processing times. These findings highlighted the reliability and practical efficiency of 3D ResNet, making it suitable for LULC applications in resource-constrained environments.

Table 6. McNemar's test results for primary dataset.

	SVM	RF	XGBoost	1D CNN	3D CNN	3D VGG	3D DenseNet	3D ResNet
SVM	-	0.134	0.248	0.023*	0.013*	0.023*	0.001*	0.004*
RF		-	0.248	0.044*	0.041*	0.617	0.041*	0.023*
XGBoost			-	0.041*	0.023*	0.074	0.001*	0.004*
1D CNN				-	0.48	0.480	0.074	0.134
3D CNN					-	0.077	0.248	0.248
3D VGG						-	0.023*	0.023*
3D DenseNet							-	1.000
3D ResNet								-

Table 7. McNemar's test results for external validation dataset.

	SVM	RF	XGBoost	1D CNN	3D CNN	3D VGG	3D DenseNet	3D ResNet
SVM	-	0.724	0.724	0.027*	0.039*	0.006*	0.003*	0.001*
RF		-	0.773	0.009*	0.027*	0.003*	0.023*	0.001*
XGBoost			-	0.027*	0.070	0.008*	0.016*	0.009*
1D CNN				-	0.617	0.617	0.617	0.617
3D CNN					-	0.480	0.480	0.248
3D VGG						-	1.000	1.000
3D DenseNet							-	1.000
3D ResNet								-

Limitations and Challenges

The superiority of 3D ResNet model comes with computational trade-offs, which may limit real-time or deployment for large-scale LULC mapping. To overcome this, transfer learning from pre-trained 2D models via DEWT technique was integrated to reduce training time. Although this approach alleviated some computational burden, the reliance on powerful hardware highlights the possible limited accessibility to powerful GPUs. These computational demands influenced model selection as more complex architectures could improve accuracy but with the cost of increased processing time and resource requirements.

Further optimization like model pruning and quantization could further reduce computational demands. While these methods were not explored in this study, they represent practical ways to balance accuracy and efficiency. Aside from that, cloud-based distributed computing could enable scalable model training across multiple devices, improving accessibility for resource-constrained environments. Techniques like knowledge distillation (Hinton et al., 2015) or mixed-precision training (Micikevicius et al., 2018) could also enhance computational efficiency. Alternative setups such as edge computing could provide a practical solution for deploying DL models in large-scale applications. Future research could explore these strategies to improve the feasibility and scalability of 3D models for LULC mapping.

Dataset imbalances were addressed partially via SKCV but it remains a concern, especially for models like SVM which are sensitive to imbalanced data. While SKCV ensures that each fold

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

represents the class distribution, this may not fully mitigate bias in model evaluation. Other technique like oversampling is a common approach in which the minority classes are replicated randomly until a balanced dataset distribution is achieved. On the other hand, undersampling reduces instances from the majority classes. As for cost-sensitive learning, this works by placing a higher misclassifying cost for the minority classes without altering the original dataset distribution. Moreover, synthetic data generation like Synthetic Minority Oversampling Technique (SMOTE) could create new samples for the minority classes, further improving model generalization. Combining these techniques with SKCV could potentially enhance model performance and reduce bias.

The lack of interpretability in DL models also poses significant challenges in domains like LULC classification. Understanding how a model arrives at its predictions is vital for enhancing interpretability (Li et al., 2022). Methods like Class Activation Mapping (CAM) and SHapley Additive exPlanations (SHAP) can aid in prediction explanation and visualizations. Future work could focus on integrating these techniques to improve transparency in DL applications for LULC classification.

Future research could also focus on utilizing deeper 3D variants of ResNet (e.g. 3D ResNet-50) architectures to enhance classification results. Deeper 3D ResNet model could potentially work more effectively than the current 3D ResNet-18 model since it will be able to capture more intricate details from the features. Although recently developed hybrid models like Vision Transformer (ResNet-ViT) is not widely adopted in remote sensing field, it also shows great potential as it can capture both local and global information from the image. Due to cost and data availability constraints, we validated our model using an existing dataset from the same satellite system but with a different location. Future work can also focus on model generalizability to a completely different geographical area or satellite dataset such as Sentinel-2. This would help with evaluation of model transferability while considering cost-effective alternatives.

4. Conclusion

In this research, the possible application of pre-trained 3D ResNet for multispectral land use and land cover (LULC) classification is highlighted by its great capability of capturing spectral-spatial features with high accuracy. The results showed that 3D ResNet outperformed other models by achieving the highest OA of 99.66% and KA of 99.39% on the primary dataset. While its performance on the external validation dataset (OA: 82.89%, KA: 80.05%) was slightly lower than that of the 3D DenseNet, it showed great efficiency with processing times of 490.2 minutes for primary dataset and only 3.6 minutes for validation dataset. McNemar's test results further showed consistent significant statistical differences (p<0.05) in classification performance between 3D ResNet and 3D DenseNet with the other models. Overall, this research contributes to the improvement of multispectral LULC classification in terms of accuracy and efficiency using advanced DL model integrated with transfer learning technique.

5. Acknowledgement

The researchers would like to thank the Ministry of Higher Education (MOHE), Malaysia for its support and resources through the Long-Term Research Grant Scheme (LRGS) of the Malaysian Research University Network (MRUN). This project is being undertaken under the research programme: 'A Big Data Analytics Platform for Optimizing Oil Palm Yield Via Breeding by Design (Grant No: 203.PKOMP.6770007)' with specific project: 'Geoinformatics Data for Palm Oil Yield Prediction Using Machine Learning (Vote No: 6300268-10801)'. Also, the authors wish to acknowledge the expertise provided by the team from FGV R&D Sdn Bhd.

6. References

Abdi, A. M. (2020). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience & Remote Sensing*, *57*(1), 1–20. https://doi.org/10.1080/15481603.2019.1650447

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

D, E., & Bhavani, N. P. G. (2023). An Effective DNN Based ResNet Approach for Satellite Image Classification. *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, 1055–1062. https://doi.org/10.1109/ICOSEC58147.2023.10276330

Ebrahimi, A., Luo, S., & Chiong, R. (2020). Introducing Transfer Learning to 3D ResNet-18 for Alzheimer's Disease Detection on MRI Images. *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1–6. https://doi.org/10.1109/IVCNZ51579.2020.9290616

Firat, H., Asker, M. E., Bayindir, M. İ., & Hanbay, D. (2023). 3D residual spatial–spectral convolution network for hyperspectral remote sensing image classification. *Neural Computing and Applications*, *35*(6), 4479–4497. https://doi.org/10.1007/s00521-022-07933-8

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network (arXiv:1503.02531). arXiv. https://doi.org/10.48550/arXiv.1503.02531

Huang, G., Liu, Z., Maaten, L. van der, & Weinberger, K. Q. (2018). *Densely Connected Convolutional Networks* (arXiv:1608.06993). arXiv. https://doi.org/10.48550/arXiv.1608.06993

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)

Jombo, S., Adam, E., Byrne, M. J., & Newete, S. W. (2020). Evaluating the capability of Worldview-2 imagery for mapping alien tree species in a heterogeneous urban environment. *Cogent Social Sciences*, *6*(1), 1754146. https://doi.org/10.1080/23311886.2020.1754146

Jozdani, S. E., Johnson, B. A., & Chen, D. (2019). Comparing Deep Neural Networks, Ensemble Classifiers, and Support Vector Machine Algorithms for Object-Based Urban Land Use/Land Cover Classification. *Remote Sensing*, *11*(14), Article 14. https://doi.org/10.3390/rs11141713

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization (arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2018). Mixed Precision Training (arXiv:1710.03740). arXiv. https://doi.org/10.48550/arXiv.1710.03740

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., Bian, J., & Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems, 64(12), 3197–3234. https://doi.org/10.1007/s10115-022-01756-8

Liu, J., Wang, T., Skidmore, A., Sun, Y., Jia, P., & Zhang, K. (2023). Integrated 1D, 2D, and 3D CNNs Enable Robust and Efficient Land Cover Classification from Hyperspectral Imagery. *Remote Sensing*, *15*(19), Article 19. https://doi.org/10.3390/rs15194797

Noh, S.-H. (2021). Performance Comparison of CNN Models Using Gradient Flow Analysis. *Informatics*, *8*(3), 53. https://doi.org/10.3390/informatics8030053

Shaharum, N. S. N., Shafri, H. Z. M., Ghani, W. A. W. A. K., Samsatli, S., Al-Habshi, M. M. A., & Yusuf, B. (2020). Oil palm mapping over Peninsular Malaysia using Google Earth Engine and machine learning algorithms. Remote Sensing Applications: Society and Environment, 17, 100287. https://doi.org/10.1016/j.rsase.2020.100287

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. https://doi.org/10.1109/JSTARS.2020.3026724

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., & Zhang, L. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, *237*, 111322. https://doi.org/10.1016/j.rse.2019.111322

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. NeuroImage, 277, 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

Vali, A., Comai, S., & Matteucci, M. (2020). Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sensing*, *12*(15), 2495. https://doi.org/10.3390/rs12152495

Wang, J., Bretz, M., Dewan, M. A. A., & Delavar, M. A. (2022). Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects. *Science of The Total Environment*, *822*, 153559. https://doi.org/10.1016/j.scitotenv.2022.153559

DOI:https//doi.org/10.22452/mjs.vol44no4.3
Malaysian Journal of Science 44(4): 18-30 (December 2025)