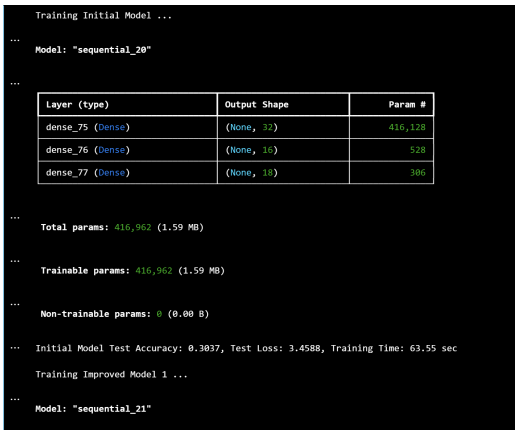
Reflection Report: Deep Learning Model Comparison for Location Classification

This project aimed to classify Myanmar location names into their respective regions using neural network models. The Initial Model consisted of 3 dense layers and a total of 416,962 trainable parameters. It served as a baseline with a validation accuracy of 0.3037 and a validation loss of 3.4588, completing training in 63.55 seconds.



Training Initial Model ...

Model: "sequential_20"

Layer (type)	Output Shape	Param #
dense_75 (Dense)	(None, 32)	416,128
dense_76 (Dense)	(None, 16)	528
dense_77 (Dense)	(None, 18)	306

Total params: 416,962 (1.59 MB)

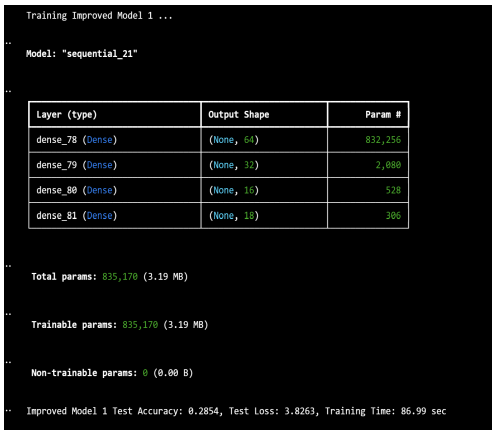
Trainable params: 416,962 (1.59 MB)

Non-trainable params: 0 (0.00 B)

Initial Model Test Accuracy: 0.3037, Test Loss: 3.4588, Training Time: 63.55 sec

Training Improved Model 1 ...

Model: "sequential_21"



Training Improved Model 1 ...

Model: "sequential_21"

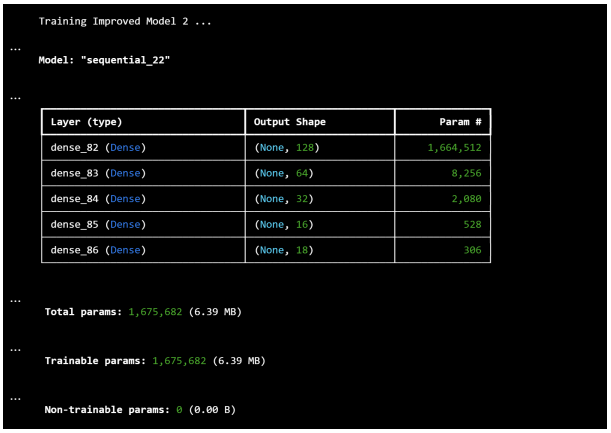
Layer (type)	Output Shape	Param #
dense_78 (Dense)	(None, 64)	832,256
dense_79 (Dense)	(None, 32)	2,080
dense_80 (Dense)	(None, 16)	528
dense_81 (Dense)	(None, 18)	306

Total params: 835,170 (3.19 MB)

Trainable params: 835,170 (3.19 MB)

Non-trainable params: 0 (0.00 B)

Improved Model 1 Test Accuracy: 0.2854, Test Loss: 3.8263, Training Time: 86.99 sec



Training Improved Model 2 ...

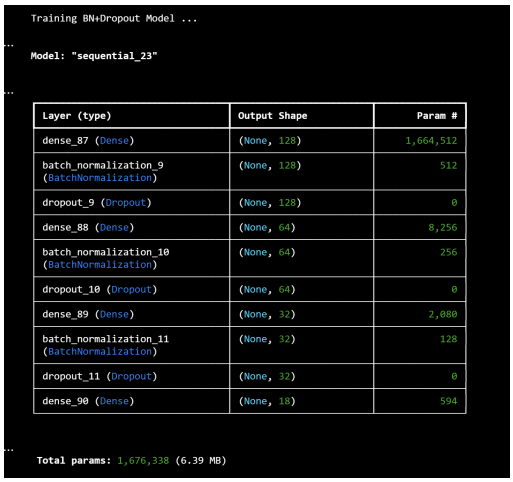
Model: "sequential_22"

Layer (type)	Output Shape	Param #
dense_82 (Dense)	(None, 128)	1,664,512
dense_83 (Dense)	(None, 64)	8,256
dense_84 (Dense)	(None, 32)	2,080
dense_85 (Dense)	(None, 16)	528
dense_86 (Dense)	(None, 18)	306

Total params: 1,675,682 (6.39 MB)

Trainable params: 1,675,682 (6.39 MB)

Non-trainable params: 0 (0.00 B)



Training BN+Dropout Model ...

Model: "sequential_23"

Layer (type)	Output Shape	Param #
dense_87 (Dense)	(None, 128)	1,664,512
batch_normalization_9 (BatchNormalization)	(None, 128)	512
dropout_9 (Dropout)	(None, 128)	0
dense_88 (Dense)	(None, 64)	8,256
batch_normalization_10 (BatchNormalization)	(None, 64)	256
dropout_10 (Dropout)	(None, 64)	0
dense_89 (Dense)	(None, 32)	2,080
batch_normalization_11 (BatchNormalization)	(None, 32)	128
dropout_11 (Dropout)	(None, 32)	0
dense_90 (Dense)	(None, 18)	594

Total params: 1,676,338 (6.39 MB)

Model Improvement Analysis:

Model	Layers	Parameters	Val Accuracy	Val Loss	Training Time(s)
Initial Model	3	416,962	0.3037	3.4588	63.5469
Improved Model 1	4	835,170	0.2854	3.8263	86.9885
Improved Model 2	5	1,675,682	0.2755	3.4624	132.3054
BN+Dropout Model	10	1,676,338	0.2614	3.7970	197.3934

Analysis and Findings:

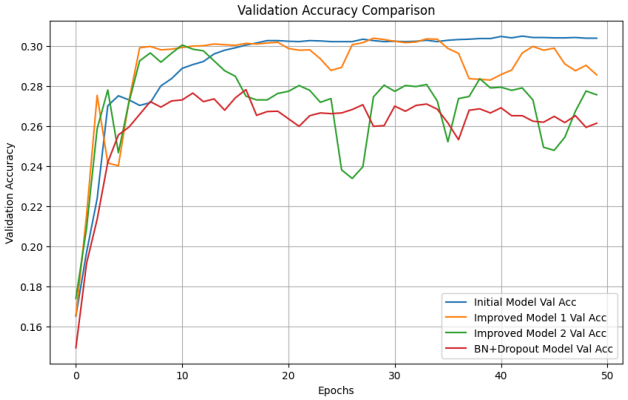


Figure 1: Validation Accuracy Comparison

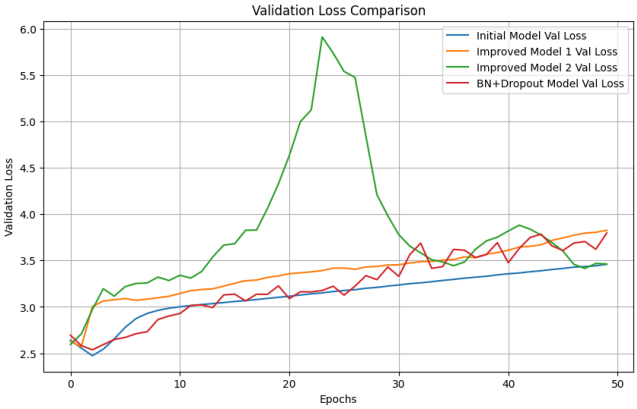


Figure 2: Validation Loss Comparison

	Model	Dataset	Precision	Recall	F1-Score
0	Initial Model	Test	0.5660	0.2242	0.2841
1	Initial Model	Train	0.8918	0.8650	0.8764
2	Improved Model 1	Test	0.5477	0.2240	0.2751
3	Improved Model 1	Train	0.8968	0.8597	0.8752
4	Improved Model 2	Test	0.5539	0.2194	0.2727
5	Improved Model 2	Train	0.8715	0.8394	0.8523
6	BN_Dropout_Model	Test	0.5024	0.2060	0.2505
7	BN_Dropout_Model	Train	0.8212	0.7838	0.7995

Confusion Matrix - Classification Model

Figure 3: Initial Model

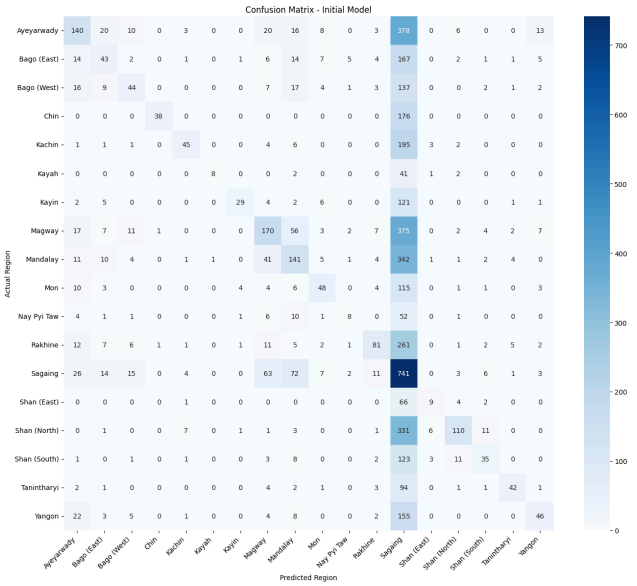


Figure 4: Improved Model 1

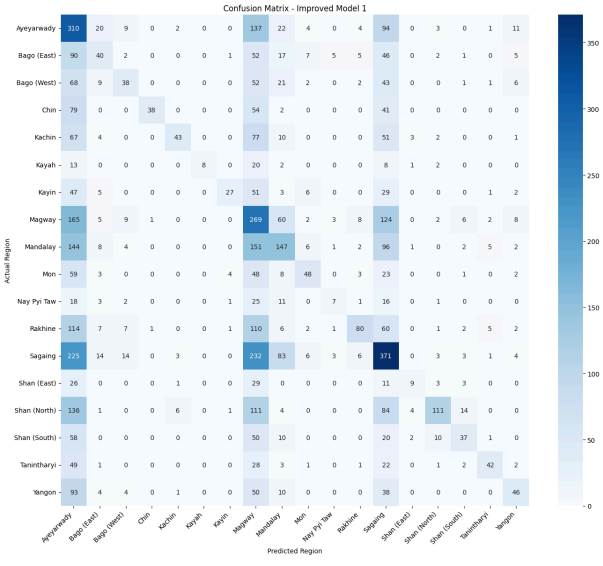


Figure 5: Improved Model 2

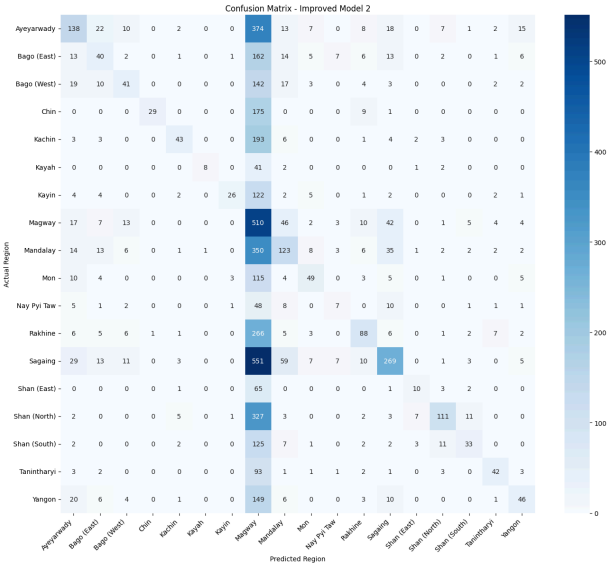
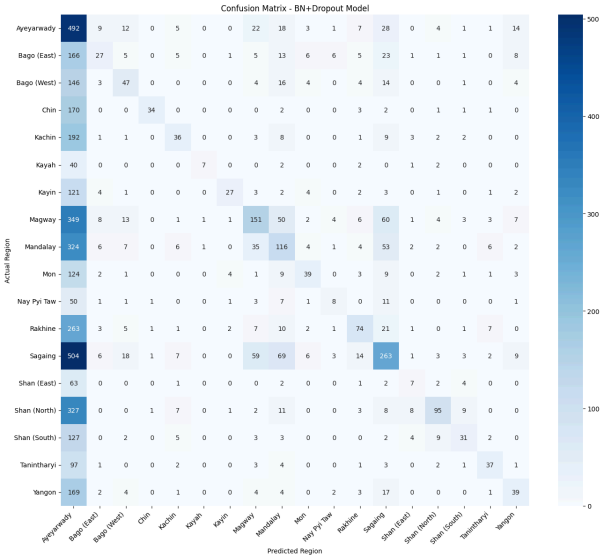


Figure 6: BN+Dropout Model



To enhance performance, several improved models were developed by **adding more layers** and increasing model depth and capacity.

Despite the increased complexity, **none of the deeper models (including the BN+ dropout model) outperformed the initial architecture**. Validation accuracy dropped slightly in all cases. Moreover, training times increased significantly with deeper architectures (up to **197.39 seconds**). F1-scores and confusion matrices confirmed that deeper models overfit to training data and fail to generalize better to unseen examples.

In conclusion, **the Initial Model delivered the best performance** in terms of accuracy, generalization, and efficiency. The added depth and regularization did not lead to improvements, likely due to the high dimensionality of sparse input features and class imbalance.

Confusion matrix analysis revealed consistent misclassifications across all models, particularly among Ayeyarwady, Sagaing, Magway, and Mandalay regions. Smaller regions like Chin, Kayah, and Nay Pyi Taw showed better classification accuracy. However, none of the improved models were able to address core region-level ambiguities.

Overall, adding more layers or introducing batch normalization and dropout did not improve test performance. The models suffered from overfitting, as seen in the gap between training F1-scores (~0.87) and test F1-scores (<0.29). This suggests that increasing complexity alone is insufficient without better input representations or more aggressive regularization. Future work should explore alternative features like TF-IDF embeddings, advanced regularization techniques, or pre-trained models to enhance generalization.