CHAPTER V

CONCLUSIONS AND SUGGESTIONS

5.1. Conclusion

Based on the results of the research that has been conducted, the following conclusions can be drawn:

1. The Cross Vision Transformer (CrossViT) method was successfully implemented to classify AI-generated images and Manual Illustrator images using a multi-scale dual-branch architecture with S-Branch (patches 16×16, 196 tokens) and L-Branch (patches 32×32, 49 tokens) connected through a cross-attention mechanism.
2. The CrossViT-15 model, which was trained using 140 trained images with ImageNet pretrained weights, managed to achieve a validation accuracy of 90.00% in the 7th epoch and a test accuracy of 93.55% in the test data consisting of 31 images.
3. The model produces a Precision value of 88.24%, a Recall of 100%, and an F1-Score of 93.75%. A 100% Recall value indicates that the model is capable of detecting all 15 AI-Generated images without any of them passing as manual images (False Negative = 0).
4. Of the 31 test data images, only 2 were misclassified, both of which were Manual Illustrator images that were predicted to be AI-Generated, which visually had a very smooth texture and consistently resembled the output of the generative model.
5. From testing five hyperparameter scenarios with ImageNet pretrained weights, Scenario 4 (High LR) produced the best performance with a Test Accuracy of 96.67% and an F1-Score of 96.77% using a combination of learning rate = 1×10^{-3}, batch size = 64, weight decay = 0.05, patch size = 16:32, and warmup epochs = 10. These results confirm that the use of ImageNet pretrained weights with more aggressive learning rate configurations and larger batch sizes can significantly improve model performance on limited datasets.

5.2. Suggestions

Based on the results of the study, some suggestions for further development are as follows:

1. The number of datasets needs to be significantly enlarged beyond the 201 images in order for the model to learn more diverse patterns and generate better generalizations, especially in images with ambiguous visual characteristics between the two classes.
2. Testing of larger CrossViT variants such as CrossViT-18 or merging with other architectures such as Swin Transformer can be performed to increase feature representation capacity and potential model accuracy.
3. Hyperparameter scenario testing should be done with a larger number of epochs and still using *pretrained weights* so that comparisons between scenarios better reflect the actual performance under optimal conditions.