ABSTRACT

Diabetic retinopathy is an eye disease that requires comprehensive retinal image analysis to accurately determine its severity level. Convolutional Neural Network (CNN)-based models still face limitations in capturing the global spatial context that is crucial for this task. Vision Transformer (ViT) offers the ability to model visual context more comprehensively through its self-attention mechanism, while Conditional Ordinal Regression for Neural Networks (CORN) models ordinal relationships between class levels more consistently through a series of binary classification tasks and conditional probabilities. This study aims to develop a diabetic retinopathy severity classification model by integrating a pretrained ViT as a global feature extractor with CORN as the ordinal prediction scheme using the RetinaMNIST dataset. The dataset consists of 1,600 images, divided into 1,080 training, 120 validation, and 400 test samples The official MedMNIST benchmark sets ResNet-18 and ResNet-50 as the best-performing baseline models. The results show that the ViT-CORN model achieves the best performance with an AUC of 0.8831; Accuracy of 0.6500; Balanced Accuracy of 0.6152; MAE of 0.4450; RMSE of 0.8093; Macro-F1 of 0.5984; and QWK of 0.8017. These results surpassed the performance of ResNet-18 and ResNet-50. The confusion matrix analysis also indicates that model errors occur among adjacent ordinal classes. These findings suggest that the ViT-CORN approach provides more accurate predictions for diabetic retinopathy severity classification.

Keywords: *Diabetic Retinopathy*, RetinaMNIST, *Ordinal Classification*, *Vision Transformer*, CORN