# Evaluating Deep Learning for Clinical Decision-Making: A Case Study on Diabetic Retinopathy

Anders Sylvest Jespersen

June 10, 2025

**Abstract**

Deep learning has shown promise in automating medical image analysis, particularly in predicting diabetic retinopathy (DR), a diabetes-related eye condition that can lead to blindness. The study Nakayama et al. (2024a) reports strong performance (AUC-ROC = 0.97 for the ConvNeXtv2 model) using two convolution neural networks (CNNs) on retinal fundus images. However, critical limitations remain in the study, such as lack of reproducibility, reliance on data leakage datasets, lack of external validation, limited model architectures, and narrow evaluation metrics.

This thesis addresses these gaps by investigating model reproducibility, data leakage, model architecture and size, data efficiency, and external validation. Recent advances in large foundation models have demonstrated strong generalization capabilities across domains, including medical imaging. In addition to CNN-based models, the study evaluates vision transformer-based (ViTs) foundation models, including DINOv2, RetFound, and VisionFM. Evaluation metrics include conventional measures like macro AUC-ROC and F1-score, alongside clinically relevant tools such as calibration curves and the Polytomous Discrimination Index (PDI), which is an extension of AUC-ROC for multi-class classification.

The results show that reproduced patient-stratified models has lower performance (AUC-ROC = 0.93 for the ConvNeXtv2 model). The larger ViT models do not outperform CNNs. Ine general, all models show poor calibration affects the reliability of the models. External validation further reveals challenges in generalizability (AUC-ROC = 0.56 for the ConvNeXtv2 model). These findings emphasize the need for rigorous, multifaceted evaluation in developing AI tools for clinical use and caution against over-interpreting results from studies that rely on limited validation or narrow evaluation criteria.

# Contents

**1 Introduction** 4

**2 Background Knowledge** 7

2.1 Diabetic Retinopathy . . . 7
2.2 Classify Diabetic Retinopathy with Deep Neural Network . . . 10
2.3 Deep Learning . . . 10
2.3.1 Transfer Learning . . . 14
2.3.2 Convolution Neural Network . . . 14
2.3.3 Vision Transformers . . . 16
2.3.4 Supervised Learning . . . 18
2.3.5 Self-supervised Learning . . . 19
2.4 Performance Metrics . . . 22
2.4.1 F1-score . . . 22
2.4.2 AUC-ROC . . . 23
2.4.3 PDI . . . 24
2.4.4 Calibration Plot . . . 24

**3 Method** 25

3.1 Use of Generative AI . . . 25
3.2 Data . . . 25
3.2.1 BRSET . . . 25
3.2.2 mBRSET . . . 27
3.3 Data Split . . . 27
3.4 Models . . . 28
3.4.1 ConvNextv2 Large . . . 28
3.4.2 ResNet-200d . . . 29
3.4.3 DINOv2 Large . . . 30
3.4.4 RetFound . . . 30
3.4.5 VisionFM . . . 30
3.5 Pre-training Data . . . 31
3.6 Fine-tuning . . . 31
3.6.1 Pre-processing . . . 31
3.6.2 Training . . . 32
3.7 Model Evaluation . . . 34
3.7.1 Ordinal PDI . . . 35
3.8 Experimental Setup . . . 36
3.8.1 Objective 1: Replication . . . 36

3.8.2 Objective 2: Data Leakage . . . . . . . . . . . . . . . . . . . . . . 36
3.8.3 Objective 3: Larger Models . . . . . . . . . . . . . . . . . . . . . . 36
3.8.4 Objective 4: Additional Performance Metrics . . . . . . . . . . . . 37
3.8.5 Objective 5: Data Efficiency . . . . . . . . . . . . . . . . . . . . . . 37
3.8.6 Objective 6: External Validation . . . . . . . . . . . . . . . . . . . 37

**4 Results & Discussion 37**
4.1 Objective 1: Replication . . . . . . . . . . . . . . . . . . . . . . . . . . 37
4.2 Objective 2: Data Leakage . . . . . . . . . . . . . . . . . . . . . . . . . 38
4.3 Objective 3: Larger Models . . . . . . . . . . . . . . . . . . . . . . . . . 40
4.4 Objective 4: Additional Performance Metrics . . . . . . . . . . . . . . . 48
4.5 Objective 5: Data Efficiency . . . . . . . . . . . . . . . . . . . . . . . . 48
4.6 Objective 6: External Validation . . . . . . . . . . . . . . . . . . . . . . 51

**5 Conclusion 56**

**6 Acknowledgment 57**

**7 Appendices 63**

**A Confusion Matrices 63**
A.1 Objective 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
A.2 Objective 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
A.3 Objective 6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65
A.4 Objective 6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67

**B Tables of Discriminative Performances 70**
B.1 Objective 5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 70
B.2 Objective 6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71

**C Calibration Plots 73**
C.1 Objective 5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
C.2 Objective 6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 75

**D Distribution of Predicted Probabilities 76**
D.1 Objective 5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
D.2 Objective 6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78