A Study of Continuous Regression for Diabetic Retinopathy Grading with Intra- and Cross-Dataset Evaluation

Mateus Bastos Magalhães Mar[1], Diana Maria Da Câmara Gorayeb[1]

[1] Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)

{mateusbastos261}@gmail.com

{dgorayeb}@uea.edu.br

Abstract. *Diabetic Retinopathy (DR) severity follows an ordinal progression, although it is often treated as a multiclass classification problem. This study presents an empirical comparison of binary, multiclass, and continuous regression-based formulations using EyePACS and APTOS 2019, under intra- and cross-dataset settings. Results show that the linear formulation achieves higher Quadratic Weighted Kappa than softmax in the evaluated scenarios, while dataset characteristics substantially affect performance. The study provides a comparative analysis of these formulations under a unified experimental protocol, rather than proposing a new ordinal learning method.*

1. Introduction

Diabetic Retinopathy (DR) is one of the leading microvascular complications of diabetes mellitus and a major preventable cause of vision loss among working-age adults [Seo et al. 2025]. Its progression is typically asymptomatic during the early stages, making regular ophthalmologic screening a critical component for preventing irreversible visual impairment [Seo et al. 2025]. In this context, retinal fundus image analysis plays a central role in the diagnosis and severity grading of the disease, a process traditionally performed by specialists following standardized clinical grading protocols.

With the advancement of deep learning, convolutional neural network (CNN)-based models have been widely explored to automate DR detection and severity grading [Seo et al. 2025]. These models, particularly architectures pre-trained on large-scale natural image datasets, have demonstrated strong capability in extracting complex visual features from medical images [Sugeno et al. 2021]. Despite these advances, the performance of such systems remains highly dependent on dataset characteristics, including image acquisition quality, class distribution, and the presence of noise [Sugeno et al. 2021].

DR severity grading is not a purely categorical problem but rather an ordinal one, consisting of five severity levels within a continuous clinical progression: 0, 1, 2, 3, and 4, representing no DR, mild, moderate, severe, and proliferative DR, respectively. In this setting, misclassifications between adjacent grades are clinically less severe than errors involving distant categories [Sugeno et al. 2021]. Although conventional multiclass formulations remain widely used, previous studies have also investigated approaches that

explicitly incorporate the ordinal structure of DR severity [Pak et al. 2020]. At the same time, variations among datasets collected in different clinical environments introduce additional generalization challenges, particularly under domain shift conditions [Chopra et al. 2025].

This work presents an empirical comparison of different modeling formulations for DR grading, including binary classification, conventional multiclass classification, and continuous regression applied to the ordinal severity scale of the disease. Rather than introducing a new ordinal learning method, the study evaluates these formulations under a common EfficientNetB2-based training framework using the EyePACS and APTOS 2019 datasets. Both intra-dataset and cross-dataset evaluations are considered to examine how the different formulations behave under distinct data distributions and dataset shift.

The comparison focuses on the relationship between output formulation, loss function, ordinal consistency, and overall predictive performance. By evaluating these alternatives under a unified experimental protocol, the study provides an empirical analysis of their relative strengths and limitations for automated DR grading, with particular emphasis on their behavior under cross-dataset conditions.

2. Literature Review

Automatic Diabetic Retinopathy (DR) grading using deep learning has achieved significant advances in recent years, particularly through the adoption of transfer learning-based architectures. Among these, the EfficientNet family has demonstrated competitive performance in DR severity grading tasks and has been widely employed on datasets such as APTOS 2019 [Sugeno et al. 2021].

Beyond architectural design, previous studies have explicitly incorporated the ordinal nature of DR into the learning process. For example, [Pak et al. 2020] showed that an ordinal encoding strategy applied to the output of an EfficientNet-B4 model outperformed conventional multiclass classification, increasing the Kappa coefficient from 0.744 to 0.790 on the APTOS 2019 dataset.

[Patni et al. 2025] investigated the impact of dataset variability on diabetic retinopathy classification by comparing different deep learning architectures across the EyePACS and APTOS 2019 datasets. Their findings indicated that model performance is strongly influenced by the characteristics of each dataset, highlighting the challenges of generalization under varying image quality conditions and class distributions.

To address these limitations, [Zhang et al. 2024] explored domain adaptation techniques aimed at reducing discrepancies between retinal imaging datasets. Their results demonstrated that adaptation strategies can improve model generalization across multiple datasets, highlighting their potential to mitigate the effects of domain shift in automated diabetic retinopathy classification.

Complementarily, the systematic review conducted by [Chopra et al. 2025] emphasized that models achieving high performance in internal validation often experience substantial degradation when evaluated on external datasets. The authors attributed this behavior to factors such as demographic differences, image acquisition devices, and domain shift, underscoring the importance of evaluating models across

multiple datasets to obtain a more realistic assessment of their generalization capabilities.

Overall, the literature highlights the relevance of ordinal modeling and cross-dataset evaluation for DR grading. This study therefore provides an empirical comparison of different output formulations under a unified experimental framework, focusing on their performance across datasets and under dataset shift.

3. Development

The methodology adopted in this study was adapted from the Cross-Industry Standard Process for Artificial Intelligence (CRISP-AI), an extension of the Cross-Industry Standard Process for Data Mining (CRISP-DM) designed for projects involving large and complex datasets, Machine Learning (ML), and Artificial Intelligence (AI). The methodology encompasses the phases of problem understanding (presented in the Introduction), data acquisition, data preprocessing, data stratification and splitting, data augmentation, deep learning architecture design, model training, evaluation metrics, experimental pipelines, and results analysis.



Figure 1. CRISP-AI-based methodology adapted for the processing and analysis of images obtained from the datasets.

The following sections provide detailed descriptions of the remaining phases of data processing and analysis.

3.1 Data Sources

The first dataset selected for this study is **EyePACS**[1], originally released as part of a Kaggle competition. It consists of retinal fundus images acquired under diverse imaging conditions and using different acquisition devices, resulting in substantial variability in image quality, illumination, and focus. The dataset contains diabetic retinopathy severity grades ranging from 0 to 4 and exhibits a highly imbalanced class distribution, with a predominance of images without retinopathy. Additionally, each image is associated with

a patient identifier and the corresponding eye laterality (left or right eye).

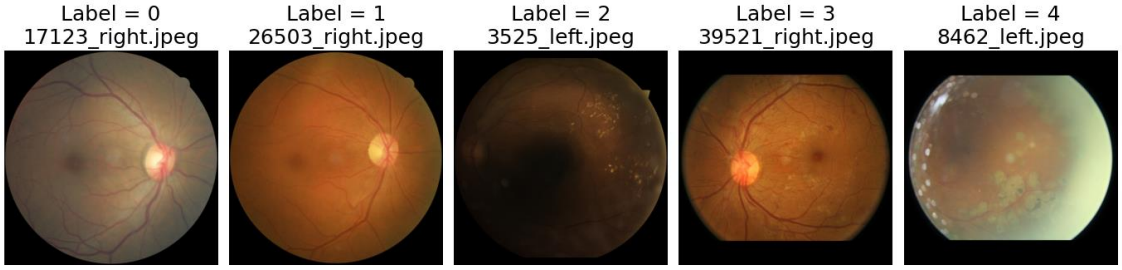


Figure 2: Representative fundus images corresponding to the five diabetic retinopathy severity levels in the EyePACS dataset.

The second dataset employed in this study is **APTOS 2019**[2], also released through a Kaggle competition. The images were acquired in a more controlled clinical environment, although variations in image quality, noise levels, and class distribution remain present across the dataset.

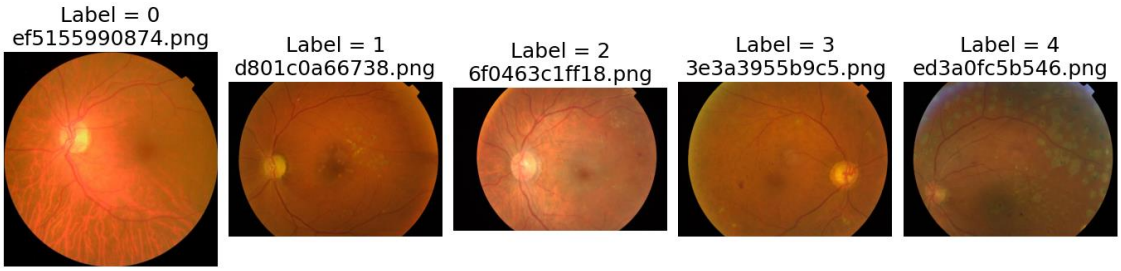


Figure 3: Representative fundus images corresponding to the five diabetic retinopathy severity levels in the APTOS 2019 dataset.

The differences between the datasets in terms of both scale and class distribution were intentionally explored to evaluate model robustness under conditions of variability and class imbalance. Table 1 summarizes the class distributions of the datasets used in this study:

Table 1. Distribution of diabetic retinopathy severity classes in the datasets used.

Dataset	Class 0	Class 1	Class 2	Class 3	Class 4	Total
EyePACS	25.810	2.443	5.292	873	708	35.126
APTOS 2019	1.805	370	999	193	295	3.662

3.2 Preprocessing

The images were first converted from the BGR color space to RGB and subsequently subjected to an automatic cropping procedure based on an intensity threshold of 7 applied to the grayscale version of each image. This procedure removed dark peripheral borders while preserving the original image whenever no valid region remained after mask application. After cropping, all images were resized to 456 × 456 pixels using INTER_AREA interpolation. Pixel value normalization was not performed during this preprocessing stage; instead, it was applied later as part of the model training pipeline.

[1]https://www.kaggle.com/c/diabetic-retinopathy-detection
[2]https://www.kaggle.com/c/aptos2019-blindness-detection

3.3 Stratification and Data Splitting

The datasets were divided into training, validation, and test sets using stratified sampling to preserve the distribution of DR severity levels. To ensure comparability between the intra-dataset and cross-dataset evaluation scenarios, sampling proportions were defined such that the datasets used in each experimental pipeline contained similar absolute numbers of images, despite the substantial differences in scale between EyePACS and APTOS 2019, while maintaining the original class proportions.

For the EyePACS experiments, a subset corresponding to 10% of the original dataset size was first sampled using stratification to preserve the original class distribution. This reduced subset was then divided into training, validation, and test sets using a stratified 60/20/20 split, preserving the class proportions within the reduced dataset. Patient-level separation was enforced to prevent data leakage between subsets. For APTOS 2019, no data reduction was applied; instead, the complete dataset was divided into training, validation, and test sets using the same stratified 60/20/20 split, preserving its original class distribution.

For the cross-dataset experiments, models were trained on one dataset and evaluated on the other. In the EyePACS-to-APTOS scenario, a subset corresponding to 10% of the original EyePACS dataset size was first sampled using stratification and then divided into training and validation sets using a stratified 80/20 split. A subset corresponding to 10% of the original APTOS 2019 dataset size was separately sampled using the same stratified procedure and used as the test set. In the reverse scenario, the complete APTOS 2019 dataset was divided into training and validation sets using a stratified 80/20 split, without data reduction, whereas a subset corresponding to 1% of the original EyePACS dataset size was sampled using stratification and used for final evaluation. Thus, the 10% and 1% reductions refer to the total number of images in the respective datasets, rather than to independent percentages assigned to each class. Stratified sampling was used to maintain the original class distribution in each reduced subset and in subsequent partitions.

Experimental reproducibility was ensured by fixing the random seed to 1234 throughout the pipeline. The seed was applied during data sampling, patient-level stratified splitting, and initialization of stochastic training components, including NumPy, TensorFlow, and data generators. Additionally, TensorFlow deterministic mode was enabled to minimize variability across experimental runs.

3.4 Data Augmentation

Data augmentation was applied exclusively to the training set to increase data variability, standardize model inputs, and improve robustness to natural variations in retinal fundus images.

Light geometric transformations were used to simulate common acquisition variations, including changes in image framing and patient positioning, while preserving diagnostically relevant anatomical structures. These transformations comprised random rotations, horizontal and vertical shifts, zooming, and horizontal flipping.

The augmentation pipeline was implemented using the ImageDataGenerator class, with rotations up to 15°, horizontal and vertical shifts up to 5%, zoom up to 10%, and horizontal flip. Pixel values were also rescaled to the [0,1] range using rescale=1/255.

The same augmentation settings were used across all experiments to ensure direct comparability between pipelines and avoid performance differences caused by augmentation strategies.

3.5 Deep Learning Architecture

The experiments were conducted using the EfficientNetB2 architecture as the deep convolutional neural network. The model was initialized with weights pre-trained on ImageNet (transfer learning), using only the convolutional backbone (`include_top=False`).

Feature representations were extracted by the backbone from the input images (RGB format with dimensions 260 × 260 × 3, resized again by the TensorFlow/Keras data generators to match the network input requirements). A Global Average Pooling layer was then applied, followed by a Dropout layer with a rate of 0.5 to reduce overfitting.

Different output configurations and loss functions were evaluated and organized into distinct experimental pipelines. In the traditional categorical approach, a dense layer with five neurons and a softmax activation function was employed, trained using the categorical cross-entropy (CCE) loss function. In the multiclass pipelines with softmax output, class weights inversely proportional to class frequencies in the training sets were applied to reduce the impact of class imbalance among DR severity levels and improve model sensitivity for underrepresented classes. A binary configuration with sigmoid activation and binary cross-entropy (BCE) loss was also evaluated.

To explore the ordinal nature of DR, a regression-based approach was implemented using a single-neuron linear output layer and the Mean Squared Error (MSE) loss function, allowing the model to learn proximity relationships between disease severity levels. MSE was selected because it quadratically penalizes errors according to the distance between severity levels, thereby emphasizing larger discrepancies along the ordinal DR scale. Compared with alternatives such as Mean Absolute Error (MAE), which applies a linear penalty and reduces sensitivity to distant errors, MSE proved more suitable for emphasizing predictions that are inconsistent with the clinical progression of the disease. Furthermore, the use of a linear output enables severity levels to be represented on an ordered scale, preserving the ordinal structure of the problem without relying on categorical encodings. For evaluation, the continuous outputs of the linear model were converted into the five severity levels by rounding each prediction to the nearest integer, with decision boundaries at 0.5, 1.5, 2.5, and 3.5. Values outside the 0–4 range were clipped to the nearest valid severity level.

This formulation maintains a consistent backbone architecture and training protocol across all approaches, enabling a direct comparison between the different modeling strategies. In all configurations, training was performed using the Adaptive Moment Estimation (Adam) optimizer to optimize learning performance.

3.6 Training

All models were implemented using the TensorFlow/Keras framework. Training was performed with a batch size of 4 and a maximum limit of 100 epochs. Due to the high image resolution and the computational cost associated with training deep convolutional neural networks, a reduced learning rate of 5×10^{-6} was adopted.

In all pipelines, the Adam optimizer was used together with a dynamic learning rate adjustment strategy (ReduceLROnPlateau), reducing the learning rate by half whenever no improvement in the monitored metric was observed for 10 consecutive epochs.

The Early Stopping technique was also employed, with restoration of the best weights obtained during training, aiming to reduce overfitting and avoid unnecessary training after convergence. In the binary pipelines, monitoring was performed using the Binary F1-score metric, calculated on the validation set at the end of each epoch, whereas in the pipelines with linear and softmax outputs, the Quadratic Weighted Kappa (QWK) metric was used. A patience value of 20 epochs was adopted for these monitoring procedures.

3.7 Experimental Pipelines

Multiple experimental pipelines were conducted to compare different strategies for DR classification, varying network output type, loss function, and training/evaluation dataset combinations. The experiments included binary, categorical multiclass, and regression-based multiclass approaches, as well as cross-dataset generalization settings. Train/validation/test splits followed the protocol described in Section 3.3.

Table 2. Summary of experimental pipelines.

Type	Output	Loss	Train	Val	Test
Binary	Sigmoid	BCE	EyePACS	EyePACS	EyePACS
Multiclass	Softmax	CCE	EyePACS	EyePACS	EyePACS
Multiclass	Linear	MSE	EyePACS	EyePACS	EyePACS
Binary	Sigmoid	BCE	APTOS	APTOS	APTOS
Multiclass	Softmax	CCE	APTOS	APTOS	APTOS
Multiclass	Linear	MSE	APTOS	APTOS	APTOS
Multiclass	Linear	MSE	EyePACS	EyePACS	APTOS
Multiclass	Linear	MSE	APTOS	APTOS	EyePACS
Multiclass	Softmax	CCE	EyePACS	EyePACS	APTOS
Multiclass	Softmax	CCE	APTOS	APTOS	EyePACS

3.8 Evaluation Metrics

The evaluation metrics were defined according to the ordinal nature of the DR classification problem, in which there is a continuous clinical progression between levels (0–4). Although metrics such as accuracy and F1-score are widely used in multiclass tasks, they treat all classes as independent categories and penalize all classification errors equally. For this reason, the **QWK** metric was prioritized as the main evaluation criterion in multiclass scenarios, as it explicitly incorporates the ordinal structure of the disease, penalizing distant errors on the scale (e.g., 0→4) more heavily than adjacent errors. In addition, QWK measures agreement between predictions and labels corrected for

agreement expected by chance, making it more suitable for assessing the clinical consistency of predictions.

In binary classification pipelines, performance monitoring was mainly based on **F1-score** and accuracy, with F1 being used as the model selection criterion for saving the best model. This choice is motivated by the potential class imbalance and the need to balance sensitivity and precision in detecting the presence versus absence of disease.

In multiclass pipelines with softmax output, the main optimization and checkpoint criterion was QWK, complemented by accuracy and F1-score (macro and weighted), enabling a more robust evaluation of both global agreement and per-class performance. In models with continuous output, QWK was also used as the primary validation metric, alongside **Mean Absolute Error (MAE)**, which provides a direct measure of average error on the ordinal disease scale, as well as F1-score and accuracy after discretization of predictions.

During training, callbacks were configured to log additional metrics at each epoch and to save the model only when an improvement in the dominant validation metric of each pipeline was observed (F1 in binary models, QWK in multiclass models). This ensures consistency between the optimization objective and the expected clinical behavior of disease progression, reducing the risk of models with high overall accuracy but low ordinal consistency.

4. Results

4.1 Overall Performance Analysis

The results obtained by the different pipelines are presented from a comparative perspective, including intra-dataset (EyePACS and APTOS 2019) and cross-dataset evaluations. Three modeling formulations are considered: binary classification, conventional multiclass classification, and continuous regression. The Quadratic Weighted Kappa (QWK) is used as the primary metric for multiclass comparison because it accounts for the ordinal distance between severity levels.

Table 3 below presents the results on the APTOS 2019 test set:

Table 3 — Results on the APTOS 2019 test set

Pipeline	Output	Loss	QWK	Binary F1	Macro F1	Weighted F1	Accuracy	MAE
Binary	Sigmoid	BCE	—	0.9728	—	—	0.9727	—
Multiclass	Softmax	CCE	0.8298	—	0.5910	0.7489	0.7462	—
Multiclass	Linear	MSE	0.8605	—	0.5411	0.7332	0.7312	0.4228

In the APTOS 2019 dataset, all multiclass models showed higher performance than those observed on EyePACS (Table 4). The linear formulation achieved the highest QWK (0.8605), while the softmax formulation obtained slightly higher Macro F1 and Accuracy values. These results indicate that the relative performance of the output formulations depends on the evaluation metric, with the linear formulation showing an advantage when ordinal agreement is emphasized.

The best overall result was obtained by the linear model with MSE loss, achieving a QWK of **0.8605**. Although the softmax model achieved slightly higher Macro F1 and Accuracy values, the linear model exhibited greater ordinal consistency, reflected in its higher QWK score.

Table 4 below presents the results on the EyePACS dataset (test set):

Table 4 — Results on the EyePACS test set

Pipeline	Output	Loss	QWK	Binary F1	Macro F1	Weighted F1	Accuracy	MAE
Binary	Sigmoid	BCE	—	0.5275	—	—	0.7044	—
Multiclass	Softmax	CCE	0.2767	—	0.2618	0.4855	0.4204	—
Multiclass	Linear	MSE	0.4252	—	0.2980	0.5978	0.5610	0.6150

In the EyePACS dataset, substantially lower performance was observed across the multiclass models compared with APTOS 2019. The linear formulation obtained the highest QWK (0.4252), compared with 0.2767 for the softmax formulation. It also achieved higher Macro F1, Weighted F1, and Accuracy. These results provide an empirical comparison in which the linear formulation performed better than the conventional softmax formulation under the evaluated EyePACS setting.

Among the multiclass models, the linear formulation based on implicit ordinal regression achieved the best overall results, reaching a QWK of **0.4252**, significantly outperforming the traditional softmax model, which obtained a QWK of **0.2767**. This result suggests that approaches that preserve ordinal continuity between diabetic retinopathy levels are better able to model the clinical progression of the disease.

The binary model achieved an accuracy of **0.7044** and an F1-score of **0.5275**. This difference illustrates the effect of reducing the five-level grading task to a binary distinction and further motivates the use of ordinal multiclass metrics when the objective is severity grading rather than disease detection alone.

Finally, Table 5 presents the cross-dataset results for APTOS and EyePACS on the test set:

Table 5 — Cross-dataset results on the test set

Train	Test	Output	Loss	QWK	Macro F1	Weighted F1	Accuracy	MAE
APTOS	EyePACS	Linear	MSE	0.4066	0.3102	0.6380	0.6524	0.5907
EyePACS	APTOS	Linear	MSE	0.7125	0.3946	0.6123	0.5956	0.6584
APTOS	EyePACS	Softmax	CCE	0.3845	0.2614	0.6229	0.6609	—
EyePACS	APTOS	Softmax	CCE	0.6765	0.2825	0.3985	0.3715	—

The cross-dataset experiments provide an empirical comparison of the Softmax and Linear formulations under domain shift conditions. In both evaluated scenarios, the linear approach outperformed the Softmax model. In the APTOS→EyePACS setting, the linear model achieved a QWK of **0.4066**, compared with **0.3845** for Softmax. Similarly, in the EyePACS→APTOS setting, the linear formulation achieved a QWK of **0.7125**, compared with **0.6765** for Softmax.

In addition to higher QWK values, the linear models also achieved better results in Macro F1 and Weighted F1 in both experimental settings. Only in the APTOS→EyePACS scenario was a small advantage observed for the Softmax model in Accuracy (**0.6609** versus **0.6524**). However, this difference was not reflected in the other metrics, where the linear approach consistently maintained superior performance. This behavior suggests that continuous modeling of disease severity produces predictions that are more consistent with the ordinal structure of diabetic retinopathy, even when the number of exact classifications is similar.

An asymmetry was observed between the two cross-dataset evaluation directions across both output formulations. For the Linear formulation, training on EyePACS and evaluating on APTOS 2019 resulted in a QWK of **0.7125**, compared with **0.4066** when training on APTOS 2019 and evaluating on EyePACS. A similar pattern was observed for the Softmax formulation, with QWK values of **0.6765** and **0.3845**, respectively. These results indicate that cross-dataset performance was substantially influenced by the direction of dataset transfer, regardless of the output formulation.

Overall, the cross-dataset results show that the linear formulation achieved higher QWK, Macro F1, and Weighted F1 than Softmax in the evaluated settings. These findings provide an empirical comparison of the two output formulations under dataset shift and indicate that their relative performance should be assessed using metrics that account for the ordinal structure of DR.

4.2 Discussion of output formulations

Previous studies have examined diabetic retinopathy performance from different perspectives, including variations in CNN architectures, transfer learning strategies, and ordinal modeling. For example, [Patni et al. 2025] compared different architectures across EyePACS and APTOS 2019, highlighting the influence of dataset variability on model performance. In contrast, the present study keeps the EfficientNetB2 backbone fixed and empirically compares different output formulations under the same training framework, including conventional softmax and continuous linear regression.

An important aspect of the observed results is the distinction between exact classification performance and ordinal error severity. The higher QWK obtained by the Linear formulation indicates greater ordinal consistency, since QWK assigns higher penalties to misclassifications involving larger distances between the true and predicted severity levels. Thus, the advantage of the Linear formulation was not necessarily restricted to achieving more exact classifications, but also involved producing errors that were comparatively closer to the correct severity level. In this sense, the results suggest that the Linear formulation was less prone to large deviations along the DR severity scale, particularly under the evaluated cross-dataset conditions.

The results can also be contrasted with [Zhang et al. 2024], who investigated source-free domain adaptation to reduce inter-domain discrepancies across retinal datasets. While their approach focuses on adapting representations between domains, the present study does not introduce a domain adaptation mechanism. Instead, it evaluates whether different output formulations exhibit different behavior when applied within and across datasets. Under the evaluated settings, the linear formulation achieved higher QWK than Softmax in both cross-dataset scenarios, providing an empirical comparison of these formulations under dataset shift.

5. Conclusion

This study presented an empirical comparison of binary, conventional multiclass, and continuous regression-based formulations for automated Diabetic Retinopathy grading using a common EfficientNetB2 framework. The experiments were conducted on the public EyePACS and APTOS 2019 datasets under both intra-dataset and cross-dataset settings.

The results showed that binary models achieved the highest accuracy in the intra-dataset experiments, reflecting the simpler nature of the binary detection task. Among the multiclass formulations, the linear output with MSE achieved the highest QWK in both intra-dataset evaluations, reaching 0.8605 on APTOS 2019 and 0.4252 on EyePACS. In the cross-dataset experiments, the linear formulation also achieved higher QWK, Macro F1, and Weighted F1 than Softmax in the evaluated settings.

The comparison also showed substantial differences between datasets. APTOS 2019 yielded higher performance across the evaluated configurations, whereas EyePACS presented lower performance, highlighting the influence of dataset characteristics on DR grading. Cross-dataset evaluation further revealed a performance degradation and an asymmetric behavior depending on the training and test datasets, illustrating the challenges associated with domain shift.

Moreover, the higher QWK achieved by the Linear formulation suggests greater ordinal consistency, with prediction errors that were comparatively less severe in terms of their distance from the true DR severity level. This indicates that the advantage of the Linear formulation was not limited to exact classification performance, but also concerned the ordinal proximity of its predictions.

Overall, the findings provide an empirical comparison of commonly used output formulations for DR grading under a unified experimental framework. The results indicate that the linear formulation was generally superior to the conventional Softmax formulation in the evaluated datasets and cross-dataset settings, particularly in terms of QWK. Rather than introducing a new ordinal learning method, this study contributes an empirical comparison of these formulations across datasets and under dataset shift.

Future work may investigate specialized ordinal methods, such as ranking losses and cumulative link models, as well as additional external datasets and domain adaptation strategies.

References

Shearer, C. The New Blueprint for Data Mining. Journal of Data Warehousing, v. 5, n. 4, p. 13-22, 2000.

Patni, K., Yagnik, S., and Patel, P. (2025). Exploring dataset variability in diabetic retinopathy classification using transfer learning approaches. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, 7(3):763–777.

Chopra, M., Sparrenberg, L., Berger, A., Khanna, S., Terheyden, J. H., and Sifa, R. (2025). From retinal pixels to patients: Evolution of deep learning research in diabetic retinopathy screening. *arXiv preprint* arXiv:2511.11065.

Zhang, G.-H., Zhuo, G.-P., Zhang, Z.-X., Sun, B., Yang, W.-H., and Zhang, S.-C. (2024). Diabetic retinopathy identification based on multi-source-free domain adaptation. *International Journal of Ophthalmology*, 17(7):1193–1204.

Sugeno, A., Ishikawa, Y., Ohshima, T., and Muramatsu, R. (2021). Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Computers in Biology and Medicine*, 137:104795.

Islam, M. R., Abdulrazak, L. F., Nahiduzzaman, M., Goni, M. O. F., Anower, M. S., Ahsan, M., Haider, J., and Kowalski, M. (2022). Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. *Computers in Biology and Medicine*, 146:105602.

Saxena, G., Verma, D. K., Paraye, A., Rajan, A., and Rawat, A. (2020). Improved and robust deep learning agent for preliminary detection of diabetic retinopathy using public datasets. *Intelligence-Based Medicine*, 3–4:100022.

Katada, Y., Ozawa, N., Masayoshi, K., Ofuji, Y., Tsubota, K., and Kurihara, T. (2020). Automatic screening for diabetic retinopathy in interracial fundus images using artificial intelligence. *Intelligence-Based Medicine*, 3–4:100024.

Pak, A., Ziyaden, A., Tukeshev, K., Jaxylykova, A., and Abdullina, D. (2020). Comparative analysis of deep learning methods of detection of diabetic retinopathy. *Cogent Engineering*, 7(1):1805144.

Chong, D. D., Das, N., and Singh, R. P. (2024). Diabetic retinopathy: Screening, prevention, and treatment. *Cleveland Clinic Journal of Medicine*, 91(8):503–510.

Seo, H., Park, S.-J., and Song, M. (2025). Diabetic retinopathy (DR): Mechanisms, current therapies, and emerging strategies. *Cells*, 14(5):376.

Ansari, P., Tabasumma, N., Snigdha, N. N., Siam, N. H., Panduru, R. V. N. R. S., Azam, S., Hannan, J. M. A., and Abdel-Wahab, Y. H. A. (2022). Diabetic retinopathy: An overview on mechanisms, pathophysiology and pharmacotherapy. *Diabetology*, 3:159–175.

Yang, Z., Tan, T.-E., Shao, Y., Wong, T. Y., and Li, X. R. (2022). Classification of diabetic retinopathy: Past, present and future. *Frontiers in Endocrinology*, 13:1079217.

Mohanty, C., Mahapatra, S., Acharya, B., Kokkoras, F., Gerogiannis, V. C., Karamitsos, I., and Kanavos, A. (2023). Using deep learning architectures for detection and classification of diabetic retinopathy. *Sensors*, 23:5726.

Rom, Y., Aviv, R., Cohen, G. Y., Friedman, Y. E., Ianchulev, T., and Dvey-Aharon, Z. (2024). *Diabetes detection from non-diabetic retinopathy fundus images using deep learning methodology*. *Heliyon*, 10:e36592.

Bhoopal, S., Rao, M., and Krishnappa, C. H. (2024). *Enhanced diabetic retinopathy detection and classification using fundus images with ResNet50 and CLAHE-GAN*. *Indonesian Journal of Electrical Engineering and Computer Science*, 35(1):366–377.