

A Machine Learning-Based Approach for Classifying Optic Nerve Images Associated with Papilledema

Rohan Bojja
rbojja23@gmail.com

ABSTRACT

Idiopathic intracranial hypertension (IIH) can present with papilledema, and delayed recognition of optic disc swelling may threaten vision. This exploratory study evaluates whether a classical machine-learning pipeline can distinguish fundus images carrying repository-provided labels of normal, papilledema, and pseudopapilledema.

The public OSF archive contained 1,369 images: 779 normal, 295 papilledema, and 295 pseudopapilledema. Images were converted to 8-bit grayscale, resized to 64 x 64 pixels, scaled to [0, 1], and flattened. A stratified image-level 80:20 split (random seed 42) produced 1,095 training and 274 test images. L2-regularized logistic regression was compared with a class-prior dummy classifier, 5-nearest neighbors, and a 300-tree random forest.

Logistic regression achieved 87.2% accuracy, macro precision 0.842, macro recall 0.844, macro F1 0.840, and macro one-vs-rest AUC 0.959. It outperformed the dummy classifier (56.9% accuracy), 5-nearest neighbors (83.9%), and random forest (84.7%). Pseudopapilledema had the lowest recall (0.712), with 11 of 59 images classified as papilledema.

The results provide a reproducible image-level benchmark, not evidence of patient-level or external clinical generalization. The archive contains no patient identifiers, so a patient-grouped split could not be performed. In addition, the provenance of the repository's papilledema folder requires caution because the source publication describes a broader optic-neuropathy group. Patient-linked, clinically adjudicated, multi-center data are needed before clinical use.

Keywords: Papilledema; Pseudopapilledema; Machine learning; Logistic regression; Fundus imaging; Optic nerve classification

INTRODUCTION

Idiopathic intracranial hypertension (IIH) is characterized by elevated intracranial pressure without an identifiable structural cause. It commonly presents with headache, visual symptoms, and pulsatile tinnitus [1-3]. Papilledema is optic disc swelling caused by increased intracranial pressure; if sustained, it can produce irreversible optic nerve injury and visual loss [2,3].

Oct 2026
Vol 12. No 3.

Evaluation for IIH combines clinical examination, neuroimaging, and lumbar puncture [1,2]. Examination of the ocular fundus helps initiate this pathway, but direct ophthalmoscopy is technically difficult and relevant fundus abnormalities may be missed [4]. Pseudopapilledema is an elevated appearance of one or both optic discs without retinal nerve fiber layer edema or raised intracranial pressure. It can resemble papilledema and lead to unnecessary imaging or lumbar puncture when misclassified.

Machine-learning systems can support more standardized assessment of fundus photographs. Large, multi-center deep-learning studies have reported strong discrimination of papilledema from normal and other abnormal optic discs, but such performance depends on clinically adjudicated labels and independent external testing [5]. Classical models remain useful as transparent computational baselines.

This study asks whether a reproducible, L2-regularized logistic regression model can classify images using the three labels supplied with a public archive. It also compares that model with three baselines and reports exact class-specific metrics, confusion-matrix counts, and one-vs-rest AUC values. The task is image-label classification and does not diagnose IIH.

METHODS

Research Design

This retrospective computational study used a supervised, three-class image-classification design. All analyses were rerun from the public data archive for this revision using one prespecified stratified holdout split.

Dataset and Labels

The archive was downloaded from OSF project 2w5ce [7]. Its data.zip file had SHA-256 digest 2e7a23b83f434ecd20a84ec8c82ce2a14f45bf468f6bbb2113a8b11076ae9683 and contained 1,369 JPEG files in three directories: Normal (n = 779), Papilledema (n = 295), and Pseudopapilledema (n = 295). No exact byte-identical duplicate files were detected. The archive provided directory labels and sequential filenames but no patient identifiers, eye identifiers, clinical metadata, or labeling protocol.

The associated source publication reports that fundus photographs were obtained at Kim's Eye Hospital with a non-mydratic AFC-330 camera and cropped around the optic nerve head [6]. It describes 779 normal images, 295 pseudopapilledema images, and 295 images in a broader optic-neuropathy group that included 22 papilledema images as well as ischemic, inflammatory, diabetic, and retinal disorders [6]. Because the public archive names the corresponding directory Papilledema but does not include case-level provenance, the present analysis treats all categories strictly as repository-provided labels rather than independently verified clinical diagnoses.

Preprocessing and Reproducibility

Each image was opened with Pillow 12.3.0, converted to 8-bit grayscale, resized to 64 x 64 pixels with bilinear interpolation, divided by 255 to scale intensities to [0, 1], and flattened into a 4,096-element feature vector. Analyses used Python 3.12, NumPy 2.5.2, SciPy 1.18.1, and scikit-learn 1.9.0. The random seed was 42.

Oct 2026

Vol 12. No 3.

Split, Models, and Evaluation

A stratified 80:20 image-level split assigned 1,095 images to training (normal 623, papilledema 236, pseudopapilledema 236) and 274 to testing (normal 156, papilledema 59, pseudopapilledema 59). Patient grouping was not possible because patient identifiers were absent. The primary classifier was multinomial logistic regression with L2 regularization, $C = 1.0$, the lbfgs solver, balanced class weights, and a maximum of 2,000 iterations. Comparators used the same features and split: a class-prior dummy classifier; 5-nearest neighbors with distance weighting; and a random forest with 300 trees, balanced class weights, random seed 42, and all available processor cores.

Performance measures were accuracy; per-class precision, recall, and F1; macro averages; confusion-matrix counts; and one-vs-rest ROC AUC calculated from predicted class probabilities. No threshold tuning or model selection used the test set.

Ethical Considerations

The analysis used a publicly available, de-identified image archive and involved no direct interaction with participants. The present work did not access medical records or patient identifiers. Ethical and consent procedures for the original image collection are described in the source publication [6].

RESULTS

The revised analysis corrected the source manuscript's dataset total and replaced the earlier figures, which were inconsistent with the stated sample size and model. Table 1 compares all four classifiers on the same 274-image test set. Logistic regression had the highest accuracy (0.872) and macro F1 (0.840), exceeding the class-prior baseline by 30.3 percentage points in accuracy.

Table 1. Test-set performance of the primary model and baseline classifiers

Model	Accuracy	Macro precision	Macro recall	Macro F1	Macro AUC
Class-prior dummy	0.569	0.190	0.333	0.242	0.500
5-nearest neighbors	0.839	0.804	0.822	0.812	0.933
Random forest	0.847	0.804	0.812	0.808	0.946
Logistic regression	0.872	0.842	0.844	0.840	0.959

Table 2 reports class-specific logistic-regression performance. The model performed best on normal images ($F1 = 0.929$). Papilledema recall was 0.898, whereas pseudopapilledema recall was 0.712. Of the 59 pseudopapilledema test images, 42 were correctly classified, 11 were classified as papilledema, and 6 were classified as normal.

Table 2. Class-specific performance of logistic regression

Class	Support	Precision	Recall	F1	One-vs-rest AUC
Normal	156	0.935	0.923	0.929	0.970
Papilledema	59	0.768	0.898	0.828	0.969
Pseudo-papilledema	59	0.824	0.712	0.764	0.938

Confusion Matrix for Logistic Regression

		Normal	Papilledema	Pseudopapilledema
True class	Normal	144	5	7
	Papilledema	4	53	2
	Pseudopapilledema	6	11	42
		Predicted class		

Figure 1. Confusion matrix for logistic regression on the 274-image stratified test set. Rows are true repository labels and columns are predicted labels.

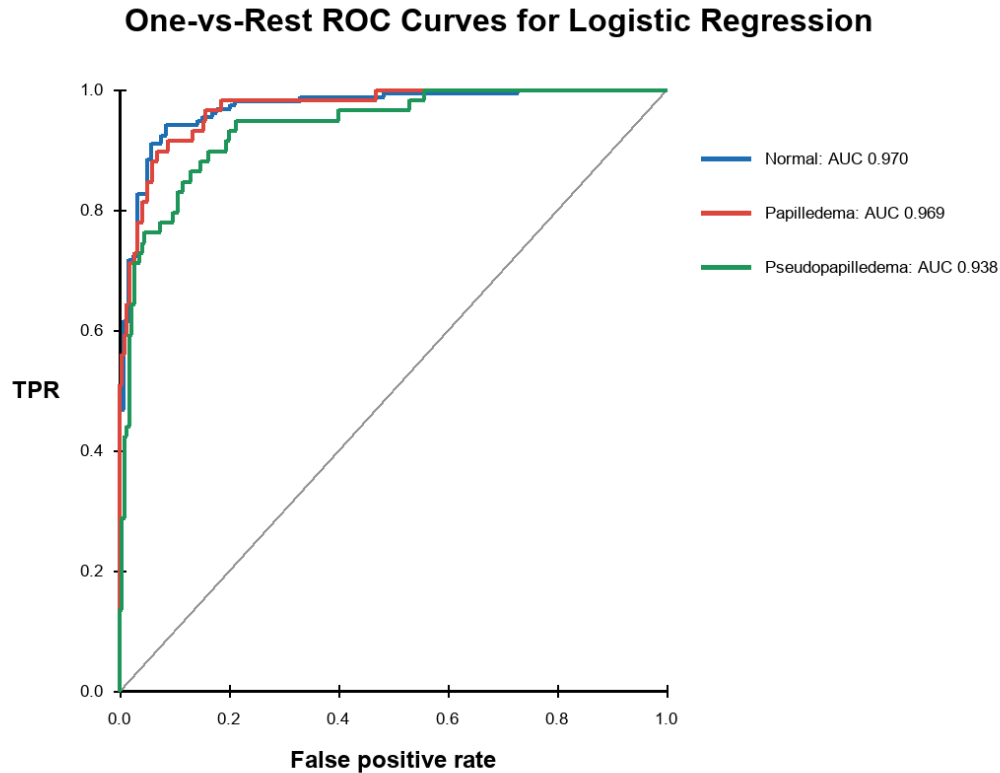


Figure 2. One-vs-rest receiver operating characteristic curves for logistic regression. AUC values were 0.970 for normal, 0.969 for papilledema, and 0.938 for pseudopapilledema.

DISCUSSION

In this reproducible image-level benchmark, logistic regression outperformed three comparators and achieved 87.2% accuracy. This result is materially lower than the previously reported 97%, which could not be reconciled with the source manuscript's stated sample size, model, and embedded figures. The revised value is therefore used throughout.

The confusion matrix provides more context than overall accuracy alone. Normal images were recognized most consistently. Papilledema recall was high, but its precision was lower because 16 non-papilledema images were assigned that label. Pseudopapilledema was the most difficult class: 11 of 59 images were classified as papilledema, consistent with overlap in optic disc appearance. These errors matter because confusing pseudopapilledema with true papilledema can prompt unnecessary diagnostic workup, whereas the reverse error could delay investigation.

The baseline comparison shows that the task contains learnable image-level signal: all three trained models exceeded the majority-class dummy classifier. Logistic regression also exceeded 5-nearest

neighbors and random forest on this split. However, differences from one holdout split should not be interpreted as definitive model rankings.

Several limitations constrain interpretation. First, the split was performed at image level because the archive has no patient identifiers. Multiple images from one patient could therefore occur in both subsets, and the reported metrics do not establish performance on unseen patients. The absence of exact duplicate files reduces one simple leakage mechanism but does not address repeated patients or correlated images. Second, all images came from one public archive associated with one clinical source and camera; there was no external validation across sites, devices, or populations. Third, the archive supplies folder labels without case-level adjudication data, and the source publication describes the nominal papilledema directory as a heterogeneous optic-neuropathy group [6]. The labels therefore should not be treated as verified diagnoses of papilledema or IIH. Fourth, resizing and grayscale conversion discard color and fine spatial information, and no confidence intervals or repeated-split uncertainty estimates were calculated.

Future work should begin with patient-linked, clinically adjudicated images and use grouped training, validation, and testing partitions. Independent multi-center testing should evaluate device and population shifts. Comparisons should also include calibrated linear models, conventional feature-based methods, and modern convolutional architectures under a shared protocol, with uncertainty intervals and prospective clinical evaluation.

CONCLUSION

L2-regularized logistic regression provided the strongest result among four evaluated classifiers on a stratified image-level split of the public archive, with 87.2% accuracy and macro one-vs-rest AUC of 0.959. The analysis is a reproducible benchmark of repository labels, not a validated screening tool for papilledema or IIH. Patient-grouped and external clinical validation are required before claims of generalization or clinical utility are warranted.

ACKNOWLEDGEMENTS

I thank Emily Broadhurst for her guidance and mentorship throughout the research process.

REFERENCES

1. Friedman DI, Liu GT, Digre KB. Revised diagnostic criteria for the pseudotumor cerebri syndrome in adults and children. *Neurology*. 2013;81(13):1159-1165. doi:10.1212/WNL.0b013e3182a55f17.
2. Mollan SP, Davies B, Silver NC, et al. Idiopathic intracranial hypertension: consensus guidelines on management. *Journal of Neurology, Neurosurgery & Psychiatry*. 2018;89(10):1088-1100. doi:10.1136/jnnp-2017-317440.

3. Wall M. Idiopathic intracranial hypertension. *Neurologic Clinics*. 2010;28(3):593-617. doi:10.1016/j.ncl.2010.03.003.
4. Bruce BB, Biousse V, Newman NJ. Nonmydriatic ocular fundus photography in neurologic emergencies. *JAMA Neurology*. 2015;72(4):455-459. doi:10.1001/jamaneurol.2014.4053.
5. Milea D, Najjar RP, Zhubo J, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *New England Journal of Medicine*. 2020;382(18):1687-1695. doi:10.1056/NEJMoa1917130.
6. Ahn JM, Kim S, Ahn KS, Cho SH, Kim US. Accuracy of machine learning for differentiation between optic neuropathies and pseudopapilledema. *BMC Ophthalmology*. 2019;19:178. doi:10.1186/s12886-019-1184-0.
7. Kim US. Machine learning for pseudopapilledema [dataset]. OSF; 2018. doi:10.17605/OSF.IO/2W5CE.
8. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.