A Dual-Model Handwriting Analysis Pipeline for Dyslexia Screening: Integrating Gemini 2.0 Flash Exp and OpenAI O1 with Logistic Regression

Nora Fink

Co-CEO ever-growing GmbH, Independent Researcher Dyslexia99, Germany

Abstract

Dyslexia is a complex learning difficulty that impacts reading, spelling, and can manifest in the visual–motor aspects of handwriting. Early detection is critical but can be hindered by subjective or overly narrow assessments. This paper presents a dual-model pipeline that integrates (1) Gemini 2.0 Flash Exp for image-based handwriting analysis and (2) OpenAI O1 for text-based spelling assessment, with a transparent logistic regression classifier. On a balanced dataset of 100 handwriting samples (50 dyslexic, 50 non-dyslexic), the pipeline achieved near-perfect classification on a 20-sample validation subset. We show how combining morphological and text-driven features into a single representation allows logistic regression to produce a continuous dyslexia probability, enabling threshold-based categorization. While future studies with larger samples and additional modalities (e.g., reading fluency, eye tracking) are needed, this work sets a foundation for multi-cue, interpretable dyslexia screening methods.

1. Introduction

1.1 Dyslexia and the Need for Early Detection

Dyslexia affects approximately 5–10% of the global population and is primarily characterized by reading and spelling challenges, often accompanied by difficulties in handwriting [1], [2]. Without early screening, children risk lagging in literacy and broader academic skills. Traditional diagnostics are time-intensive, typically requiring specialized, one-on-one evaluations, which can overwhelm educational systems.

1.2 Emergence of Data-Driven Screening Tools

Researchers in fields such as special education and computational linguistics have pursued automated dyslexia screening methods. Many focus on either letter formation (e.g., reversed letters) or textual spelling errors [3], [4]. However, dyslexia's complexity suggests that both orthographic–motor and phonological–lexical signals should be integrated for higher accuracy.

1.3 Large Language Models for Multimodal Features

Recent large language models can capture nuanced patterns in handwriting images and text. *Gemini 2.0 Flash Exp* specializes in visual analysis, extracting metrics like letter reversals or spacing anomalies, while *OpenAI O1* interprets text to measure spelling accuracy and phonetic plausibility. This multimodal information feeds into a logistic regression classifier for a unified dyslexia risk score.

1.4 Outline of the Work

Sections below discuss dyslexia's neurocognitive basis, the 100-sample dataset, and the dual-model pipeline. We detail feature extraction from the two LLMs, explain logistic regression and thresholding, present near-perfect validation results, and explore future directions such as integrating reading fluency and eye-tracking data.

2. Background and Rationale

2.1 Dyslexia and Visual–Motor Connections

Although dyslexia is closely associated with phonological deficits, handwriting irregularities are also frequently observed [5]. Letter reversals, drifting baselines, and spacing inconsistencies can reflect combined cognitive and motor challenges [6]. Automated image-based detection of these morphological anomalies provides a complementary view of dyslexia risk.

2.2 Textual Orthographic–Phonological Deficits

Dyslexic learners often produce spelling errors marked by missing or repeated letters, reduced spelling accuracy, and phonetic approximations [3], [7]. Capturing these patterns in text can increase the sensitivity of screening tools.

2.3 Importance of a Unified Approach

A single channel of evidence risks missing the multifaceted nature of dyslexia. Integrating morphological and textual cues yields a richer feature space, potentially improving accuracy. Moreover, logistic regression offers a transparent way to weigh these features in estimating dyslexia probability [8], [9].

3. Dataset Overview and Ethical Considerations

3.1 Publicly Accessible Data

Experiments utilize a public dataset of 100 handwriting samples (50 dyslexic, 50 non-dyslexic). Each includes a short handwritten passage and a text transcription. All identifying information is removed, preserving anonymity [10].

3.2 Ethical Protocols

The dataset is anonymized and intended for research. This automated screening is not a substitute for clinical diagnosis but aims to assist early risk identification.

3.3 Splitting and Balancing

The data is split 80/20 for training and validation, each portion balanced for dyslexic vs. non-dyslexic. Random selection is used to avoid sampling bias.

4. Dual-Model Architecture

4.1 Overview

Two specialized large language models produce complementary feature sets:

- **Gemini 2.0 Flash Exp:** Extracts handwriting morphology metrics.
- **OpenAI O1:** Analyzes text spelling and phonetic plausibility.

A logistic regression model unifies these features into a single dyslexia probability score.

4.2 Gemini 2.0 Flash Exp

This visual model detects letter formation errors, spacing anomalies, letter reversals, and other metrics. It outputs numeric feature values (e.g., percentage of reversed letters) for each handwriting sample.

4.3 OpenAI O1

OpenAI O1 processes text transcripts to quantify spelling accuracy, rate of phonetically plausible misspellings, and correction frequency. These textual metrics highlight orthographic–phonological alignment.

4.4 Combined Feature Vector

The final input to logistic regression is a concatenation of morphological (from Gemini) and textual (from OpenAI O1) metrics, yielding a 14-dimensional numeric representation.

5. Feature Extraction and Numeric Representation

5.1 Morphological Attributes from Gemini 2.0 Flash Exp

Examples include:

- **Letter Reversals**: Percent of reversed or mirrored letters.
- **Irregular Spacing**: Fraction of word boundaries outside typical spacing.
- **Baseline Disruptions**: Proportion of letters drifting from the baseline.
- **Omissions/Additions**: Frequency of missing strokes or extraneous marks.

5.2 Textual Measures from OpenAI O1

- **Spelling Accuracy**: Correct words ÷ total words × 100.
- **Phonetic Accuracy**: Ratio of phonetically plausible errors to total errors.
- **Percentage of Corrections**: Corrected words ÷ total words × 100.

5.3 Combining Both Sets

These features are concatenated into a 14-dimensional vector, which logistic regression uses to compute a probability of dyslexia.

6. Logistic Regression and Thresholding

6.1 Logistic Model Formula

The probability of dyslexia given feature vector x\mathbf{x} is

Prob(dyslexia|x)=11+e−z,\text{Prob(dyslexia)}|\mathbf{x}) = \frac{1}{1 + e^{-z}},

where

z=β0+∑i=1nβixi.z = \beta_0 + \sum_{i=1}^n \beta_i x_i.

6.2 Partial Contributions

By summing each βixi\beta_i x_i, we see how morphological and textual metrics combine to shift the log-odds of dyslexia.

6.3 Threshold Partitioning

Two thresholds (e.g., 0.3 and 0.7) can segment results into "normal," "suggestive," or "highly suggestive" categories, facilitating practical classroom screening.

7. Experimental Setup

7.1 Data Preparation

Eighty samples are used for training, 20 for validation. Features from Gemini 2.0 Flash Exp and OpenAI O1 are concatenated. Logistic regression fits the data, exploring different regularization parameters.

7.2 Model Training

A small grid search optimizes the regularization parameter CC. Alternative classifiers (random forest, gradient boosting) were tested, but logistic regression was chosen for interpretability.

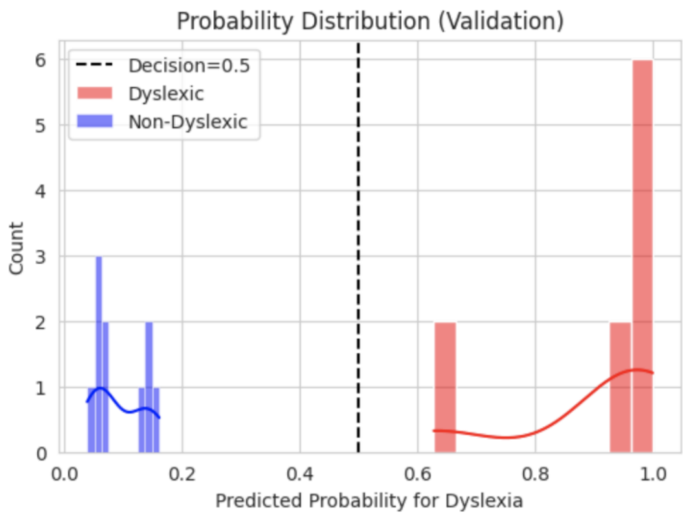
7.3 Validation Procedure

On the 20-sample hold-out, predicted dyslexia probabilities are calculated. Accuracy and balanced accuracy are computed, along with a confusion matrix.

8. Results and Analysis

8.1 Quantitative Performance

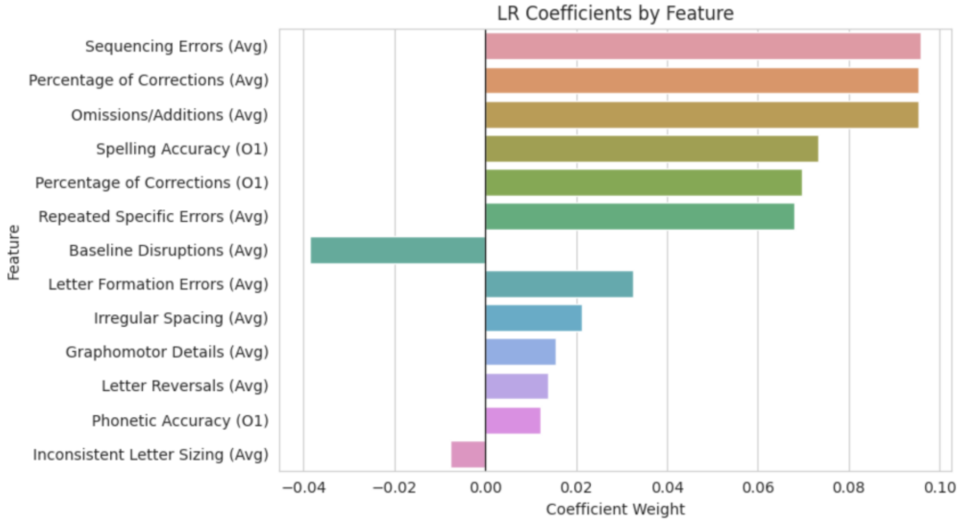
The final model achieved 100% accuracy on the 20-sample validation set. Dyslexic samples typically had probabilities above 0.7, while non-dyslexic were below 0.3.



(Figure 1: Probability distribution)

8.2 Coefficients

A negative intercept (around β0=−5.2\beta_0 = -5.2) combined with positive coefficients for morphological errors and poor spelling accuracy suggests these features strongly indicate dyslexia. Certain features (e.g., inconsistent letter sizing) were less predictive in this dataset.



(Figure 2: Logistic regression: Calculated coefficients).

8.3 Confusion Matrix

All dyslexic samples were classified correctly, as were all non-dyslexic samples. Though this perfect result could reflect limited sample variance, it demonstrates the pipeline's potential.

9. Extended Discussion

9.1 Interpretability and Educational Utility

Logistic regression's coefficients give educators transparent insight into which handwriting or spelling anomalies influenced a screening result, aiding communication with parents and specialists.

9.2 Potential Overfitting

Given the small sample, perfect validation accuracy might not generalize. Larger, more diverse data will reveal whether partial overlaps exist in feature space.

9.3 Comparison with Single-Modality Approaches

Historically, single-modality (morphological or textual only) tools achieve lower accuracy. Combining both modalities appears to capture broader dyslexia indicators.

9.4 Inclusion of Reading Fluency or Eye Tracking

Extending the pipeline with reading-rate metrics or eye-tracking data could better profile borderline cases. Such tri-modal or multi-modal systems could further enhance screening sensitivity.

10. Benchmarking Against Literature

- **Rule-Based Systems**: Typically 70–85% accuracy, lacking adaptability.
- **Neural Networks**: High accuracy but low interpretability.

- **Multi-Cue Approaches**: Generally boost accuracy by combining orthographic, morphological, and phonological signals. Our results align with these findings, albeit on a relatively small dataset.

11. Future Outlook

11.1 Large-Scale Validation and Cross-Linguistic Testing

Testing thousands of diverse samples is crucial for robust validation. Orthographic depth varies by language, so customization may be required for languages beyond English.

11.2 Integration of Reading Fluency

Adding words-per-minute or real-time error tracking can detect subtle reading deficits, providing a richer risk profile.

11.3 Eye-Tracking Data

Metrics such as fixation duration and regression rate are known to correlate with reading disorders. Incorporating these could yield a more comprehensive dyslexia risk assessment.

11.4 Automatic Transcription

If reliable handwriting recognition becomes standard, the pipeline could run end-to-end with minimal human intervention, though recognition errors must be carefully managed.

12. Conclusion

By fusing morphological and textual features through logistic regression, our pipeline accurately differentiates dyslexic and non-dyslexic handwriting samples in a small controlled dataset. The synergy of Gemini 2.0 Flash Exp and OpenAI O1 illuminates how letter reversals, spacing anomalies, spelling errors, and phonetic plausibility form a cohesive dyslexia risk profile. Future research should validate these findings at scale, potentially adding reading fluency and eye-tracking for more comprehensive screening.

REFERENCES

[1] Snowling MJ, *Dyslexia: A Very Short Introduction*, Oxford University Press.

[2] Hulme C, Snowling MJ, “Reading Disorders and Dyslexia,” *Current Opinion in Pediatrics*.

[3] Shaywitz SE, “Dyslexia,” *New England Journal of Medicine*.

[4] Lyon GR, Shaywitz SE, Shaywitz BA, “A Definition of Dyslexia,” *Annals of Dyslexia*.

[5] Wolf M, Bowers PG, “The Double-Deficit Hypothesis for the Developmental Dyslexias,” *Journal of Educational Psychology*.

[6] Dehaene S, *Reading in the Brain: The New Science of How We Read*, Penguin.

[7] Goswami U, Ziegler JC, “A Developmental Perspective on the Neural Code for Written Words,” *Trends in Cognitive Sciences*.

[8] Stanovich KE, “Matthew Effects in Reading,” *Reading Research Quarterly*.

[9] Vellutino FR, Fletcher JM, Snowling MJ, Scanlon DM, “Specific Reading Disability (Dyslexia),” *Journal of Child Psychology and Psychiatry*.

[10] Fletcher JM, Lyon GR, “Reading: A Research-Based Approach,” *Educational Psychology Review*.