Normative Modeling of Tablet Signals for Calibrated Pediatric Dysgraphia Decision Support

Shabnam Sodagari
Computer Engineering & Computer Science
California State University
Long Beach, USA
shabnam@csulb.edu

Shadnaz Asgari
Computer Engineering & Computer Science
California State University
Long Beach, USA
shadnaz.asgari@csulb.edu

Abstract

Dysgraphia can negatively affect children's education by making handwriting difficult. Binary classification of dysgraphia from tablet handwriting signals limits clinical interpretability and gives little guidance for borderline cases. We propose an age-conditional normative modeling framework that converts digital handwriting features into a continuous severity score and a Mondrian conformal triage output. We fit twelve Gaussian-process normative models on typically developing children only and define the Handwriting Maturity Gap (HMG) as each child's standardized deviation from age-typical handwriting. On 280 primary-school children, HMG alone strongly separated dysgraphic from typically developing children (AUC= 0.907) and added substantial information beyond chronological age. Within dysgraphic children, higher HMG tracked worse BHK handwriting quality, supporting its use as a severity score. A two-feature classifier using only age and HMG achieved AUC= 0.925, comparable to five 12-feature baselines. The Mondrian conformal triage layer achieved empirical coverage of 0.942, issued confident singleton decisions for 86.5% of children, and was correct on 93.3% of those singleton cases.

CCS Concepts

• **Applied computing** → **Health care information systems**; • **Computing methodologies** → *Machine learning*.

Keywords

Dysgraphia, normative modeling, Gaussian process, handwriting analysis, conformal prediction, clinical decision support, pediatric screening

ACM Reference Format:
Shabnam Sodagari and Shadnaz Asgari. 2026. Normative Modeling of Tablet Signals for Calibrated Pediatric Dysgraphia Decision Support. In *International Conference on Information Technology for Social Good (GoodIT '26), September 02–04, 2026, Pisa, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3794786.3830736

1 Introduction

Dysgraphia is a developmental handwriting disorder often linked to motor-coordination difficulties; related developmental coordination problems affect about 5% of children and may persist into adulthood [5]. Children with dysgraphia write more slowly than their peers, and their letter quality is poorer. The academic gap widens through primary school.

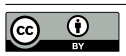
A common reference test is the *Concise Evaluation Scale for Children's Handwriting* (BHK) [5, 12]. The BHK is reliable but resource-intensive: it requires a trained specialist for administration and scoring. Specialist shortages are a cause of delayed diagnosis.

Digital tablets with smart pens make automatic screening possible. A child writes on the tablet, and software extracts dynamic features such as writing speed, pen pressure, and pen tilt. Many groups have built classifiers that predict dysgraphia from these features [2, 3, 8, 18]. Most of this work, however, treats screening as a binary problem. Nevertheless, there are two clinical needs that binary screening leaves open: (1) A binary output does not tell a clinician how severe the case is. It does not indicate which treatment to start with, or how often to follow up. (2) A dysgraphia score of 0.51 and a score of 0.99 should not lead to the same clinical action, but a binary classifier treats them the same once the threshold is crossed. The borderline cases are exactly the cases where a clinician's input matters most. Existing classifiers do not communicate this uncertainty.

Kunhoth *et al.* [15] extended dysgraphia assessment beyond binary classification by introducing the first multimodal severity dataset, comprising online and offline handwriting data from 113 Slovak children, and classifier-fusion methods for three-level grading: typically developing, grade 1 (mild), and grade 2 (moderate to significant). Their work is complementary to ours, with the main distinctions arising from three methodological and empirical differences.

Discrete versus continuous severity. Kunhoth *et al.* collapse five clinical severity levels into three label-defined classes. In contrast, our severity index is continuous: thresholds can be imposed post hoc, without quantizing the underlying estimate. This preserves information and supports longitudinal monitoring, where small within-subject changes may be clinically relevant.

Discriminative vs. generative learning. A discriminative classifier learns where the boundary between labeled groups sits. A generative model learns what one group looks like and treats deviation from it as the signal. Their classifiers are discriminative: they are trained on labeled severity classes, i.e., TD (typically developing) / grade 1 / grade 2). Our normative model is generative on the TD population only. Severity then emerges as a deviation from the age-conditional norm rather than as a learned decision boundary. This generative framing has two practical consequences: it sidesteps class-imbalance issues during normative fitting (the model never



This work is licensed under a Creative Commons Attribution 4.0 International License.
GoodIT '26, Pisa, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2483-1/2026/09
https://doi.org/10.1145/3794786.3830736

sees a dysgraphic example), and it makes the score age-conditional by construction.

Point predictions vs. calibrated sets. Their outputs are point classifications with cross-validation variance. Ours are conformal prediction sets. Split-conformal carries finite-sample marginal-coverage guarantees under exchangeability between calibration and test data; in our cross-validated evaluation, the guarantee is empirical (averaged over folds) rather than the guarantee of a single deployed model with a fixed held-out calibration set. The Mondrian variant we use stratifies by school grade, so empirical coverage is reported per grade. This adds an explicit triage layer that a binary or multiclass output cannot provide.

In other words, Kunhoth et al. [15] ask "which severity bin does this child fall into?" We ask "how far from age-typical is this child, and how confidently can we screen them?"

The contributions of this paper thus include:

(1) We define the *Handwriting Maturity Gap* (HMG), a continuous, age-conditional severity score derived from a diagonal-covariance Gaussian-process (GP) normative model (twelve independent "age to feature" GP regressions) trained on TD (typically developing) children only. HMG combines deviations across all 12 handwriting features into a single non-negative scalar.
(2) We define the *Handwriting Age Equivalent* (HAE), the age at which a child's handwriting features are most consistent with typical development. The difference between chronological age and HAE provides the clinically intuitive "writes like an N-year-old" framing.
(3) We wrap a two-feature classifier (chronological age and HMG) in Mondrian split-conformal prediction at $\alpha = 0.10$. The output is a prediction *set*: a single class for confident cases, both classes for borderline cases. Empirical cross-validated coverage is reported within each school grade.
(4) We compare against five 12-feature baselines, including an FT-Transformer.

We train and evaluate our framework on a dataset of 280 French primary school children analyzed by Gargot et al. [9]. The novelty of our method is in the *representation* (HMG, HAE) and the *decision interface* (conformal sets). We show that no statistically significant AUC difference is detected against five 12-feature baselines, and we offer interpretability and triage on top.

2 Related Work

Binary dysgraphia screening: Asselborn *et al.* [2, 3] trained a random forest on **53** handwriting features extracted from $n = 298$ children (56 dysgraphic) and achieved high diagnostic accuracy. We use the reduced 12-feature analysis subset of $n = 280$ children that Gargot et al. [9] adopted (see §3.1). Mekyska et al. [18] took a similar handwriting-feature approach on a cohort. Drotar and Dobes [8] introduced an open Slovak dataset for binary classification and reported up to 80% accuracy with AdaBoost.

Gargot et al. [9] used the same French cohort to identify *subtypes* of dysgraphia by unsupervised K-means clustering. They found subtypes that differed in motor and spatial deficits. Their contribution is orthogonal to ours: they characterized variation *within* dysgraphia, while we quantify severity *across* both groups using a continuous score grounded in typical development.

Severity grading: Kunhoth et al. [15] used a multimodal dataset and a severity-classifier benchmark. We discussed the relationship to our approach in Section 1 in brief, their target is discrete class assignment from labeled severity classes, whereas ours is a continuous deviation score from a normative model.

Direct numerical comparison between our paper and theirs is not appropriate, because the datasets (French $n = 280$, ages 6–11 vs. Slovak $n = 113$, ages 8–15), the labeling schemes (binary vs. three-class), and the metrics (ROC-AUC vs. macro-recall) all differ. We position our work as complementary: a continuous, age-conditional severity score and a calibrated screening interface, suitable for use alongside or instead of a discrete class label.

Normative modeling: Normative modeling was popularized in clinical neuroscience by Marquand et al. [17]. The approach fits a regression model on healthy controls only. Every feature value of every patient is then treated as a deviation from the age-conditional norm. Rutherford et al. [20] scaled the approach to brain imaging on 58,000 individuals. The framework has been used for schizophrenia and bipolar disorder [24]. We apply the same logic to handwriting and aggregate the per-feature deviations into a single severity score.

Conformal prediction in healthcare: Conformal prediction [1, 23] is a model-agnostic method for uncertainty quantification. Rather than returning only a single prediction, it returns a set or interval of plausible outcomes that contains the true value with a user-specified probability, without assuming a particular data distribution or model form. The Mondrian variant calibrates separately within predefined groups, such as school grade, targeting subgroup coverage under exchangeability [6, 23]. Conformal methods have also been used in medical settings, including image classification [16] and medical uncertainty [14]. To our knowledge they have not been applied to handwriting-based screening.

Tabular deep learning: Whether deep neural networks outperform gradient-boosted trees on small tabular datasets remains debated. Grinsztajn *et al.* [11] found that tree-based models still perform best on most small-data benchmarks. We therefore include FT-Transformer [10], a Transformer-based tabular model, as a neural baseline.

3 Methods

3.1 Dataset and features

The dataset by Gargot *et al.* [9] contains digitizing-tablet handwriting recordings from 280 French primary-school children performing the five-minute Concise Evaluation Scale for Children's Handwriting (BHK) copy task, the French reference test for dysgraphia. The dataset includes 218 typically developing children (TD) and 62 children with dysgraphia (DYS) across grades CP, CE1, CE2, CM1, and CM2, corresponding approximately to ages 6–10 years. Handwriting was recorded on paper affixed to Wacom Intuos tablets at 200 Hz with 0.25 mm spatial resolution. Each record includes an anonymized identifier, group, grade, age, gender, and handedness; the cohort comprises 128 female and 152 male participants, including 247 right-handed and 33 left-handed children.

The dataset contains two BHK reference scores. `BHK_raw_speed_score` records the number of characters written in five minutes, whereas

`BHK_raw_quality_score` is a handwriting-quality degradation score, with higher values indicating poorer handwriting quality. The data file reports these raw BHK scores as mean values across three raters.

We use the 12 aggregated digital handwriting features retained in Gargot et al. [9]. The static features are `Space_Between_Words`, measuring average spacing between words; `std_Density`, measuring variability in handwriting density across spatial grid cells; and `median_Freq_tremolo`, measuring the median frequency of tremor-related handwriting fluctuations. The kinematic features are `median_Freq_speed`, the median frequency of the writing-speed spectrum; `dist_Freq_speed`, the distance between the child's speed-frequency distribution and the age-matched typical distribution; and `in_Air`, the proportion of time the pen is not touching the tablet. The pressure features are `mean_Pressure`, average pen pressure; `mean_d_P`, the mean rate of pressure change; and `std_d_P`, the variability in the rate of pressure change. The pen-tilt features are `dist_Freq_tilt_x`, the distance between the child's tilt-x frequency distribution and the age-matched typical distribution; `bandwidth_tilt_x`, the bandwidth of the tilt-x frequency spectrum; and `median_Freq_tilt_y`, the median frequency of the tilt-y spectrum. Age is parsed into months from the original age strings and treated as a continuous covariate. The BHK raw quality score is used only as an external validation signal and is not included as a model input.

3.2 The Handwriting Maturity Gap

The intuition is straightforward. We learn what age-typical handwriting looks like in TD children, and we measure how far each child departs from that pattern. A larger departure means a more severe deviation from the age norm.

Why a Gaussian process? A Gaussian process (GP) regression [19] is a curve-fitting method that returns not just a predicted mean at every input, but also a predicted variance. Concretely, after training on TD data, we can ask the GP: "at age 9 years, what is the typical writing-speed value and how spread out is the GP's prediction?" and we get back both numbers. This matters here because the predictive band naturally widens at ages where TD data are sparse: the GP's posterior predictive variance is age-dependent even when the residual-noise term is constant. A z-score divided by a constant variance would be miscalibrated. A z-score divided by an age-dependent posterior predictive standard deviation $\hat{\sigma}_j(\text{age})$ accounts for this.

For each of the 12 features $j \in \{1, \ldots, 12\}$, we fit an independent GP regression on the TD-only training data. Let a denote age in months, and let $y_j(a)$ denote the value of handwriting feature j at age a. The generative model is feature-as-function-of-age plus homoskedastic Gaussian observation noise:

$$y_j(a) \;=\; \mu_j(a) \;+\; \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}\Big(0,\, \sigma^2_{n,j}\Big), \tag{1}$$

where $\mu_j(a)$ is the age-dependent mean function for feature j, ε_j is residual observation noise, and $\sigma^2_{n,j}$ is the feature-specific noise variance.

We use the covariance function

$$k_j(a, a') = \kappa_j \exp\left(-\frac{(a-a')^2}{2\ell_j^2}\right) + \sigma^2_{n,j}\delta(a, a'), \tag{2}$$

where a and a' are two ages, κ_j controls the marginal variance of feature j, ℓ_j is the length-scale controlling the smoothness of the age trend, and $\delta(a, a')$ is one when $a = a'$ and zero otherwise. This kernel is a constant times a radial basis function plus a white-noise term. The radial basis function says that data points at nearby ages should have similar feature values. The white-noise term lets the model account for residual measurement noise; we do not fit a heteroskedastic noise model.

We fit the kernel hyperparameters κ_j, ℓ_j, and $\sigma^2_{n,j}$ by marginal-likelihood maximization with two random restarts, using the implementation in scikit-learn (`RBF + WhiteKernel`). The 12 GPs are fit independently; we therefore call this a *diagonal-covariance* normative model.

The age-dependent quantity used in the z-score below is the *posterior predictive standard deviation* $\hat{\sigma}_j(a)$. This quantity reflects both uncertainty in the age-dependent feature value and the fitted homoskedastic noise variance $\sigma^2_{n,j}$. Its age dependence arises from the GP posterior and therefore tends to be larger in age regions where TD training data are sparse.

Using these age-specific GP predictions, we define the deviation z-score for child i and feature j as

$$z_{ij} \;=\; \frac{y_{ij} - \hat{\mu}_j(a_i)}{\max\big(\hat{\sigma}_j(a_i),\, 10^{-3}\big)}, \tag{3}$$

The numerator is the residual relative to the age-specific TD posterior mean $\hat{\mu}_j(a_i)$, and the denominator standardizes this residual by the corresponding posterior predictive standard deviation. The denominator is floored at 10^{-3} to avoid numerical instability when the predictive variance is very small; this floor was rarely active in practice.

The HMG score for child i is the root-mean-square z-deviation across all 12 features:

$$\text{HMG}(i) \;=\; \sqrt{\frac{1}{12}\sum_{j=1}^{12} z_{ij}^2}. \tag{4}$$

HMG is non-negative. Larger values mean larger overall deviation from the TD age-typical pattern. It is defined as the root-mean-square standardized residual across the 12 handwriting features. The diagonal approximation treats features as conditionally independent given age, which improves stability in our modest sample; A full covariance estimate would have $\binom{12}{2} = 66$ off-diagonal entries to estimate from $\sim$ 200 TD training children, which is statistically fragile.

Why predict features from age, not the reverse? A forward model predicting age from handwriting features is intuitive, but in small samples its predictions tend to shrink toward the training-set mean: young children appear too old and older children too young. We observed this pattern in pilot experiments. We therefore use the inverse formulation: predict each feature from age, then measure deviation from the TD age-typical pattern. This avoids using potentially biased age predictions as the basis for the handwriting score and follows clinical normative modeling practice [17, 20].

What HMG should look like under correct calibration. HMG is the root-mean-square of the 12 standardized residuals. If the model is well calibrated on TD children, these residuals should have mean zero and unit variance, so a typical TD child should have HMG

close to 1, not 0. Therefore, we check calibration by asking whether the average TD HMG remains close to 1 and stable across grades, rather than increasing or decreasing with age.

3.3 The Handwriting Age Equivalent

The HMG score is statistically powerful but not in clinical units. A clinician asking "What does HMG = 2.4 mean?" does not receive a simple clinical interpretation. We therefore complement HMG with the *Handwriting Age Equivalent* (HAE).

The intuition is to take a child's measured handwriting features. Ask the TD GPs "at what age would these features be most typical?" That age is the HAE. Formally, HAE is the age $\hat{a}$ that maximizes the joint Gaussian likelihood of the child's features under the per-feature TD GPs. The HAE for child i, denoted $\hat{a}(i)$, is the age at which the child's observed handwriting features are most likely under the TD GP models. Under the diagonal feature approximation, we define

$$\hat{a}(i) \;=\; \arg\max_{a} \sum_{j=1}^{12} \log \mathcal{N}\Big(y_{ij} \,\Big|\, \hat{\mu}_j(a),\ \hat{\sigma}_j^2(a)\Big), \tag{5}$$

where y_{ij} is child i's observed value for feature j, $\hat{\mu}_j(a)$ and $\hat{\sigma}_j(a)$ are the GP posterior predictive mean and standard deviation for feature j at trial age a, and $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2.

The objective is a sum of per-feature log-Gaussian densities evaluated under the TD GP models at trial age a. Because $\hat{\mu}_j(a)$ and $\hat{\sigma}_j(a)$ are nonlinear functions of age, the summed log-likelihood need not be concave or unimodal, so we do not assume an analytical optimum. We solve by exhaustive grid search over $a \in \{60, 61, \ldots, 144\}$ months, followed by three-point parabolic interpolation around the discrete maximum. This requires evaluating the 12 feature-specific GP models over 85 candidate ages per child. HAE values that fall at or near the grid boundary are limited by the $[60, 144]$-month search range and should be interpreted as bounded estimates rather than unconstrained maximum-likelihood optima.

The corresponding *Handwriting Age Gap* for child i is $\mathrm{HAG}(i) = \mathrm{age}_i - \hat{a}(i)$. A positive HAG means the child writes "younger" than their chronological age. HAE is the score we present to clinicians; HMG is the score we use for screening.

3.4 The proposed two-feature classifier

The screening classifier is $L2$-regularized logistic regression on two inputs: chronological age (z-scored using training-fold statistics) and HMG (z-scored the same way). We use class-balanced sample weights to compensate for the dys/TD imbalance. We refer to this classifier as "LogReg-AGE+HMG".

Two features may seem too few. Section 5-3 reports a comparison against five 12-feature baselines and concludes that, on this cohort, two features perform comparably to twelve.

3.5 Mondrian split-conformal prediction

We wrap the classifier in split-conformal prediction [23] so that the output is a calibrated set rather than a point prediction. Conformal prediction turns a probabilistic classifier into one that says "the answer is class c" or "the answer is one of $\{c_1, c_2\}$ — I am not sure," and gives a finite-sample guarantee on how often the true class is in the returned set.

The split-conformal procedure: Inside each outer training fold, half of the data is used to fit the classifier $\hat{f}$ and half is reserved for calibration. For each calibration sample $(\mathbf{x}_k, c_k)$, where $\mathbf{x}_k$ is the classifier input vector and $c_k \in \{\mathrm{TD}, \mathrm{DYS}\}$ is the class label, the non-conformity score is

$$s_k \;=\; 1 - \hat{p}_{\hat{f}}(c_k \mid \mathbf{x}_k), \tag{6}$$

where $\hat{p}_{\hat{f}}(c \mid \mathbf{x})$ is the probability assigned by classifier $\hat{f}$ to class c for input vector $\mathbf{x}$. These non-conformity scores, computed over all calibration samples, are the calibration scores used to choose the conformal threshold. A small s_k means that the classifier assigned high probability to the true calibration label, whereas a large s_k means that it assigned low probability to the true label. For a target miscoverage rate α, let n_{cal} be the number of calibration samples. We sort the calibration non-conformity scores and set $\hat{q}$ to the $\lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$-th smallest score. The $n_{\mathrm{cal}} + 1$ term accounts for the future test point in the rank calculation, yielding the standard split-conformal coverage guarantee under exchangeability. The prediction set for a test input $\mathbf{x}'$ is

$$\hat{C}(\mathbf{x}') \;=\; \big\{c \in \{\mathrm{TD}, \mathrm{DYS}\} \;:\; 1 - \hat{p}_{\hat{f}}(c \mid \mathbf{x}') \le \hat{q}\big\}. \tag{7}$$

In words, the set includes every class whose non-conformity score for the test input is no larger than the calibration threshold $\hat{q}$. By the standard conformal coverage theorem [1, 23], if the test point $(\mathbf{X}, G)$ and the calibration points are exchangeable, then the marginal coverage satisfies $\Pr\{G \in \hat{C}(\mathbf{X})\} \ge 1 - \alpha$.

The Mondrian variant. Coverage at $1 - \alpha$ *on average* can hide subgroups where coverage is much lower. Boström and Johansson [6] address this by computing $\hat{q}_m$ separately for each Mondrian group m, using only the calibration points with that subgroup label. This restores coverage within every subgroup. We use school grade as the Mondrian group, because the deployment workflow is grade-specific (a school nurse screens children one grade at a time). If a calibration cell has fewer than 10 children (only relevant in grade CP, which contains a single dysgraphic child in the cohort), we fall back to the pooled quantile for that fold to avoid pathological behavior on tiny calibration sets. This fallback is a deliberate trade-off: it gives up the strict per-grade Mondrian coverage guarantee for grade CP in exchange for stability, and we treat CP coverage as descriptive rather than guaranteed in the results below.

We fixed $\alpha = 0.10$, a 90% coverage target. The output is one of three sets: $\{\textsc{td}\}$ (confident TD), $\{\textsc{dys}\}$ (confident dysgraphic), or $\{\textsc{td}, \textsc{dys}\}$. The third set is a *deferral*: the model is uncertain and asks for clinician review.

3.6 Proposed analysis plan

Our goal is to verify our following hypotheses H1-H4:

H1 (HMG validity). HMG is valid if it separates dysgraphic from TD children and adds information beyond chronological age. We require HMG-only AUC > 0.75 and a likelihood-ratio test showing that `dys ~ chrono + HMG` improves over `dys ~ chrono` at $p < 0.001$.

H2 (severity gradient). If HMG captures severity beyond the binary dysgraphic/TD distinction, then HMG should correlate positively with handwriting impairment *within* the dysgraphic subgroup. We therefore test whether HMG is positively associated with the BHK raw quality score among dysgraphic children only: Spearman $\rho > 0$ with p-value $p < 0.05$.

H3 (parsimony / non-inferiority). The two-feature classifier is considered parsimonious if it performs no worse than the 12-feature baselines within a pre-specified AUC margin. Let $\Delta\text{AUC} = \text{AUC}_2 - \text{AUC}_{baseline}$, where AUC_2 is the AUC of LogReg-AGE+HMG and $\text{AUC}_{baseline}$ is the AUC of a 12-feature baseline. H3 is met if, for every baseline, $\Delta\text{AUC} > -0.02$, and at least 4 of 5 Holm-adjusted paired DeLong tests are non-significant ($p_{\text{Holm}} \geq 0.05$).

H4 (conformal triage). We consider conformal triage successful if the cross-validated marginal coverage is between 0.85 and 0.95 and the singleton rate is at least 0.70. This is an empirical calibration check, not a formal conformal guarantee. Formal split-conformal guarantees apply to one fixed deployed model with a held-out exchangeable calibration set; our 10×5 cross-validation evaluation and the grade-CP fallback to pooled quantiles do not strictly satisfy that setting.

4 Experimental Setup

Cross-validation All results use 10 repeats × 5-fold stratified cross-validation, for 50 outer folds total. Splits are stratified on the dys/TD label. The repeated protocol follows [7], reducing the variance of the mean-fold estimator at the cost of correlated folds. Inside each outer training fold, the inverse normative GPs are fit only on TD children from that fold, so no outer-test data enter the GP fits. The training fold is then split 50/50 into classifier-training and conformal-calibration subsets, while preserving the outcome-by-grade distribution as much as possible, e.g., TD-CP and dys-CP children are balanced separately where feasible.

Baselines We compare LogReg-AGE+HMG against five 12-feature baselines: (1) **LogReg-12.** *L*2-regularized logistic regression on the 12 features with class-balanced weights. (2) **RandomForest-12.** A class-balanced Random Forest using the 12 retained digital handwriting features. This baseline is inspired by the Random Forest diagnostic model of Asselborn et al. [3], adapted to the 12-feature set retained by Gargot et al. [9]. (3) **XGBoost-12.** 400 trees, depth 4, learning rate 0.05. (4) **LightGBM-12.** 400 trees, 15 leaves, learning rate 0.05. (5) **FT-Transformer-12.** 3 transformer blocks with 192-dimensional feature embeddings, 8 attention heads, 80 epochs with early stopping, AdamW optimizer, weight decay 10^{-5}, learning rate 10^{-4}, and batch size 32.

We also report two HMG-augmented baselines: **LogReg-12+AGE+HMG** (all 12 features plus chronological age plus HMG) and **XGB-12+AGE+HMG**.

Statistical tests Pairwise AUC comparisons are performed using the fast DeLong test [21]. This test compares two AUCs evaluated on the same subjects and tests the null hypothesis that the two underlying AUCs are equal. For the AUC comparison family, we use the Holm step-down correction [13], which controls the family-wise error rate. For the within-grade effect-size comparisons, we use the Benjamini–Hochberg correction [4], which controls the false discovery rate and is suitable for screening multiple subgroups. Bootstrap percentile 95% confidence intervals are computed using 2,000 resamples; bootstrap samples missing either class are discarded.

Unit of analysis for statistical tests: Repeated cross-validation produces multiple out-of-fold predictions for each child. Before computing any hypothesis test, we average these repeated predictions into one score per child. Thus, all statistical tests use the child as the unit of analysis ($n = 280$), rather than treating the repeated predictions as independent observations. This avoids inflating the effective sample size because repeated predictions for the same child are correlated.

5 Results

5.1 H1: HMG carries information beyond chronological age

HMG strongly separates dysgraphic from TD children. A two-sample Kolmogorov–Smirnov test comparing the per-child out-of-fold HMG scores gives $D = 0.77$, $p = 1.3 \times 10^{-29}$, and HMG alone achieves AUC = 0.91.

The key question is whether HMG adds information beyond chronological age. A logistic regression predicting dys from chronological age alone achieves AUC = 0.58, only slightly above chance. Adding HMG raises the AUC to 0.93. A likelihood-ratio test comparing the nested models rejects the age-only null model, $\chi^2_1 = 164.1$, $p = 1.4 \times 10^{-37}$ (Table 1). HMG was also positively associated with BHK handwriting-quality degradation, both overall and within the dysgraphic subgroup. Thus, H1 passes both decision criteria.

Table 1: HMG validation summary. All numbers are out-of-fold from 10×5 stratified cross-validation.

Quantity	Value
HMG-only AUC	0.907
Kolmogorov–Smirnov D (dys vs. TD)	0.77
Kolmogorov–Smirnov p	1.3×10^{-29}
Likelihood-ratio statistic, age vs. age + HMG	164.1
Likelihood-ratio p, age vs. age + HMG	1.4×10^{-37}
AUC, chronological-age only	0.584
AUC, chronological age + HMG	0.931
Spearman ρ (HMG, BHK quality), all	+0.42
Spearman p, all	2.6×10^{-13}
Spearman ρ, dys subgroup only	+0.32
Spearman p, dys subgroup only	0.012

Figure 1 shows the HMG distribution by group, both pooled and split by school grade. The dys distribution shifts clearly to the right in every grade. The TD HMG distribution is roughly flat across grades, with means in the range 0.89–1.07 — close to the value of 1 expected when z_{ij}^2 has unit mean under correct calibration. This stable TD pattern across grades is consistent with the inverse normative formulation avoiding the systematic age-dependent bias that we observed in the pilot forward-direction model.

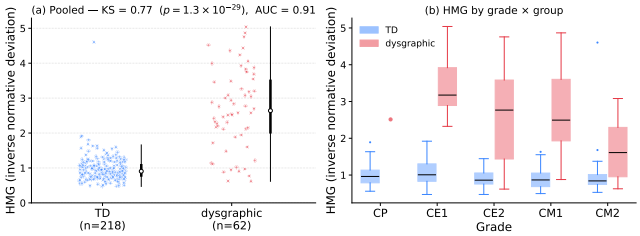


Figure 1: HMG distribution by group. (a) Pooled across all 280 children; points are individual children. (b) HMG by school grade and group; boxes show the interquartile range (IQR), whiskers are 1.5×IQR. CP contains only one dysgraphic child and is shown as a single point. The TD median is approximately stable across grades, consistent with the inverse normative model avoiding the regression-to-mean bias seen in pilot forward-direction fits.

5.2 H2: HMG captures severity within dysgraphic children

The Spearman correlation of HMG with the BHK raw quality score is $\rho = +0.42$ ($p = 2.6 \times 10^{-13}$) across all children. A between-group correlation, however, could be driven entirely by the binary class distinction: any score that perfectly separated the groups would yield a high ρ even if it carried no within-group information. A more stringent test is the within-dys correlation, which avoids the between-group separation effect by computing the correlation only inside the dysgraphic subgroup. We find $\rho = +0.32$, $p = 0.012$ (Fig. 2). The H2 decision rule is met. The within-dys correlation is modest, and BHK and HMG both encode handwriting quality, so the result *supports* rather than *proves* that HMG carries a continuous severity gradient on top of the binary class label.

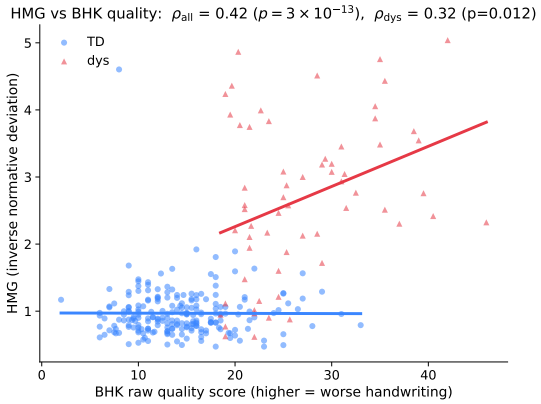


Figure 2: HMG versus BHK raw quality score. Higher BHK quality scores mean worse handwriting. The within-dys regression line is consistent with HMG carrying severity-graded information inside the dysgraphic group, on top of the binary class label.

The within-grade Cohen's d between dys and TD on HMG is large for every non-CP grade (Table 2): CE1 $d = 4.76$, CE2 $d = 2.46$, CM1 $d = 2.76$, CM2 $d = 1.17$. (Cohen's d expresses a between-group mean difference in pooled standard-deviation units; values above 0.8 are considered large effects.) All Mann–Whitney p values survive Benjamini–Hochberg correction.

Table 2: Within-grade dys-vs-TD effect sizes on HMG. Mann–Whitney p values are Benjamini–Hochberg corrected at FDR 0.05.

Grade	n_{dys}	n_{TD}	Cohen's d	p	p_{BH}
CE1	16	42	+4.76	2.7×10^{-9}	5.3×10^{-9}
CE2	15	36	+2.46	1.1×10^{-5}	1.4×10^{-5}
CM1	20	44	+2.76	1.4×10^{-9}	5.3×10^{-9}
CM2	10	48	+1.17	6.9×10^{-3}	6.9×10^{-3}

5.3 H3: Two features perform comparably to twelve

The pre-specified parsimony hypothesis is the boldest of the four hypotheses. We predicted that the two-feature LogReg-AGE+HMG classifier would match the 12-feature baselines on AUC. The intuition is that HMG already aggregates all 12 features into a single age-normalized scalar, so adding the raw features back should provide little extra signal.

Table 3 reports the result. The reference classifier achieves AUC= 0.925. Every one of the five paired DeLong tests returns a non-significant p for the null of equal AUC, with Holm-adjusted $p = 1.0$. The numerical ΔAUC (two-feature minus baseline) ranges from +0.004 (FT-Transformer) to +0.022 (LightGBM); in all five cases, the sign favors the two-feature reference classifier.

Table 3: Paired DeLong tests comparing LogReg-AGE+HMG ($AUC_2 = 0.925$) with 12-feature baselines. $\Delta AUC = AUC_2 - AUC_{baseline}$; positive values favor the two-feature model. Holm correction was applied across five tests. H3 is met: all $\Delta AUC > -0.02$ and all Holm-adjusted $p \geq 0.05$.

Baseline	AUC	ΔAUC	DeLong p	Holm p
FT-Transformer-12	0.921	+0.004	0.83	1.00
LogReg-12	0.918	+0.008	0.78	1.00
RF-12	0.914	+0.011	0.46	1.00
XGB-12	0.912	+0.013	0.53	1.00
LightGBM-12	0.904	+0.022	0.28	1.00

Reading the result: H3 is met. The decision rule required non-inferiority of LogReg-AGE+HMG to each 12-feature baseline within 0.02 AUC, with at least 4 of 5 Holm-adjusted DeLong tests non-significant. All five ΔAUC values (two-feature minus baseline) are positive, from +0.004 to +0.022, and all Holm-adjusted DeLong tests are non-significant. Thus, no baseline shows a statistically distinguishable or practically relevant ROC-AUC advantage over the two-feature model.

This supports parsimony, not formal equivalence: DeLong's null is equal AUC, whereas formal non-inferiority would require a one-sided paired confidence interval or bootstrap test against the 0.02 margin.

Figure 3 shows overlapping ROC curves for all eight classifiers, consistent with the DeLong results.

Precision–recall (PR)-AUC was reported descriptively because DeLong's test does not apply to PR-AUC. FT-Transformer-12 had the highest PR-AUC (0.887), ahead of RF-12 (0.882), XGB-12 (0.877), and LogReg-AGE+HMG (0.859). This suggests a small ranking advantage for FT-Transformer on the positive class, but does not change the H3 conclusion.

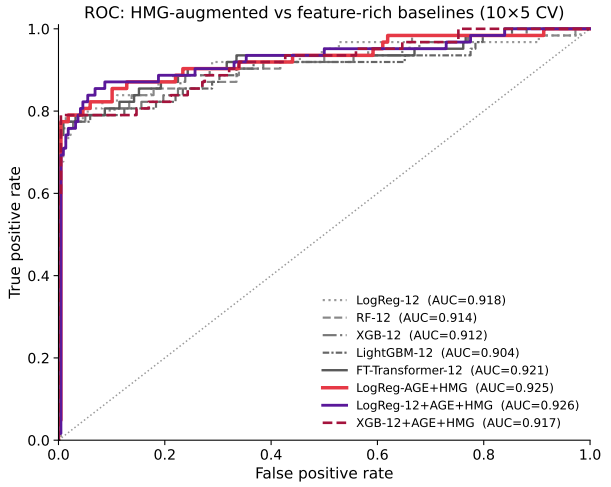


Figure 3: ROC curves for all classifiers (10×5 cross-validation, out-of-fold predictions averaged across repeats). The two-feature LogReg-AGE+HMG (red) lies on or above all five baselines. None of the pairwise differences reach statistical significance under DeLong's test (Table 3).

5.4 H4: Conformal triage meets the empirical calibration criteria

Table 4 summarizes the conformal screening operating characteristics. Figure 4 breaks them down by grade. Empirical cross-validated marginal coverage is 0.942, inside the pre-specified $[0.85, 0.95]$ interval. The singleton rate is 0.865, which exceeds the 0.70 threshold. Both H4 clauses are met. Consistent with Section 3.5, these results are interpreted as empirical cross-validated operating characteristics rather than formal split-conformal deployment guarantees. Grade CP used the pooled-quantile fallback, so its grade-specific coverage is reported descriptively.

The most clinically informative number is the *singleton accuracy*: when the conformal module returns a single class, it is correct 93.3% of the time on cross-validated out-of-fold predictions. The 13.5% deferral rate concentrates exactly where the HMG signal is weakest. Grade CM2 has the smallest within-grade Cohen's d (1.17) and also the highest deferral rate (39%). The pattern is consistent with the conformal mechanism producing larger prediction sets where the data warrant it.

5.5 Decision-curve analysis (descriptive)

Discrimination metrics like AUC and conformal coverage do not directly answer the clinical question *would deploying this classifier*

Table 4: Conformal screening operating characteristics ($\alpha = 0.10$, Mondrian by grade, 10×5 cross-validation). Coverage and singleton rates are empirical CV averages, not formal split-conformal deployment estimates.

Quantity	Value
Empirical marginal coverage (target 0.90)	0.942
Singleton rate (target ≥ 0.70)	0.865
Singleton accuracy	0.933
Deferral rate	0.135
Mean prediction-set size	1.135

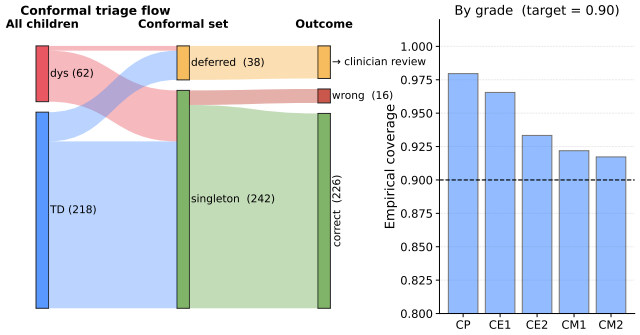


Figure 4: Conformal triage flow at $\alpha = 0.10$. Left: all 280 children grouped by ground-truth class (TD vs. dysgraphic). Middle: the conformal module's output split into *singleton* sets ({TD} or {DYS}, green) versus *deferred* sets ({TD, DYS}, amber). Right: among singletons, classifier accuracy split into correct (green) and wrong (red); deferred cases pass through to clinician review. Counts are averages across the 10 cross-validation repeats. The inset shows empirical conditional coverage by school grade against the $1 - \alpha = 0.90$ target; grade CP uses the pooled-quantile fallback (calibration cell smaller than the $n = 10$ threshold) and its CP-conditional coverage is descriptive only.

improve referral decisions versus the simple defaults of referring everyone or referring no one? Decision curve analysis [22] addresses this by plotting net benefit against the threshold probability p_t at which a clinician would be willing to refer. Net benefit is defined as $\mathrm{NB}(p_t) = \mathrm{TP}/N - \mathrm{FP}/N \cdot p_t/(1 - p_t)$ and accommodates the asymmetric costs of false positives and false negatives implicit in any chosen threshold. We computed decision curve analysis descriptively on the same out-of-fold predictions used for H1–H3.

Figure 5 shows the result. Across the clinically plausible threshold range $p_t \in [0.05, 0.5]$, the two-feature LogReg-AGE+HMG classifier dominates both *treat all* and *treat none* reference strategies; the highest-AUC 12-feature baseline tracks the two-feature curve closely, with no consistent advantage for either. The dominance over *treat all* grows rapidly as p_t exceeds the empirical prevalence (here $62/280 \approx 0.22$), which is the regime relevant to a screening setting where the clinician wants to reserve specialist evaluation for higher-probability cases. We treat this as supporting evidence

for clinical utility, not as a deployment estimate; net benefit on a held-out external cohort would be needed for the latter.

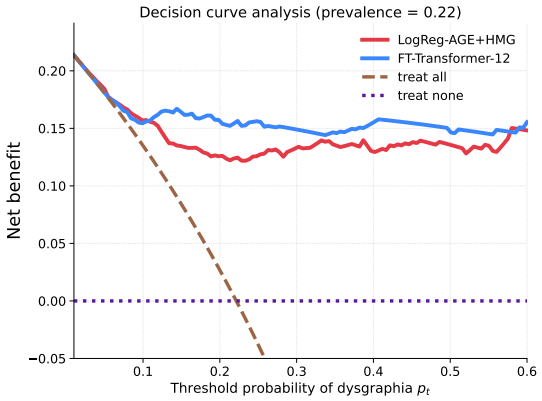


Figure 5: Decision curve analysis on 10×5 cross-validated out-of-fold predictions. The two-feature LogReg-AGE+HMG classifier (red) and the highest-AUC 12-feature baseline (blue) are plotted against the *treat all* (dashed) and *treat none* (dotted) reference strategies. Both classifier curves dominate the references across the clinically plausible threshold range. The empirical prevalence (≈ 0.22) is shown in the title.

5.6 Handwriting Age Equivalent (descriptive)

Figure 6 shows HAE versus chronological age. TD children fall on the diagonal, with mean HAG = 0.3 months, consistent with the inverse normative GP model being well-calibrated on the TD population it was trained on. Dysgraphic children fall systematically below the diagonal: their mean HAG is +11.6 months, meaning that on average a dysgraphic child writes like a TD child who is almost a year younger. Used as a score on its own, HAE achieves AUC= 0.64, lower than the diagonal-covariance HMG.

Grid-boundary sanity check. HAE is estimated as the maximum of a one-dimensional likelihood over a bounded age grid ($a \in [60, 144]$ months). Children whose features are most consistent with ages outside this range therefore receive a boundary HAE rather than an unconstrained maximum-likelihood estimate.

We quantified this by reporting the per-repeat boundary-hit rate and the proportion of children who reached a boundary in at least one repeat. The grid boundary was reached at a mean per-repeat rate of 10.6%. Across repeats, 81/280 children (28.9%) were at the boundary at least once: 37/218 TD children (17.0%) and 44/62 dysgraphic children (71.0%). Boundary hits occurred at both ends of the grid: among ever-boundary children, 27 had mean HAE < 80 months and 14 had mean HAE > 130 months; the remaining 40 hit a boundary in some repeats but landed mid-grid in others.

The higher boundary-hit rate among dysgraphic children is informative: it identifies atypical feature profiles that fall outside the normative age range modeled here. Boundary-binding HAE values are therefore best read as censored developmental-age summaries rather than unconstrained maximum-likelihood estimates. This reinforces the intended role of HAE as an interpretable clinical communication measure, while HMG remains the primary screening score.

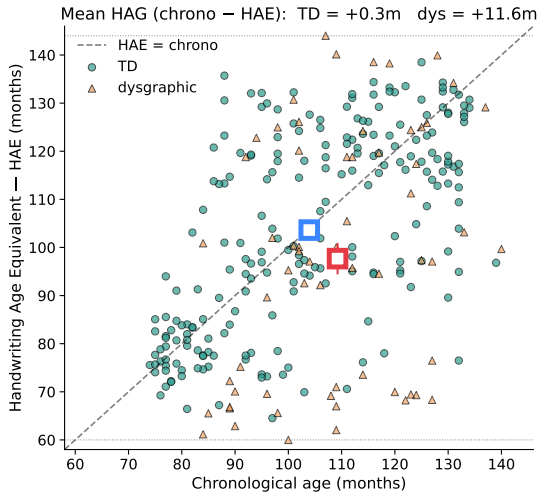


Figure 6: Handwriting Age Equivalent (HAE) versus chronological age. The dashed line is HAE = chronological age (perfect calibration). Dotted horizontal lines mark the HAE grid boundaries at 60 and 144 months (see §5.6, "Grid-boundary sanity check"). Open squares with error bars are group means with standard errors. TD children sit on the diagonal; dysgraphic children sit below it by an average of 11.6 months.

5.7 Per-feature normative trajectories

Figure 7 shows the per-feature GP normative trajectories underlying HMG. Each panel plots one of the 12 handwriting features against age: TD training points (green dots), the GP posterior mean and $\pm 1.96\sigma$ predictive band (blue), and dysgraphic children (red crosses). These GPs were fit on the full TD population for visualization only and were not used in cross-validated analyses. The figure shows what HMG measures: each child's standardized vertical deviation from the age-conditional TD mean, scaled by the GP posterior standard deviation. Several features (in-air time, tilt-x frequency, ΔP mean, density std, tremolo frequency) show a systematic upward shift of dysgraphic children above the upper TD predictive band, including in older grades where TD values continue to mature. These deviations are first aggregated into HMG and then used by the conformal module to decide whether to issue a single-class prediction or defer.

We also report deferral rates by sex and laterality. Males ($n = 152$) are deferred at 11.8%; females ($n = 128$) at 15.5%. Right-handed children ($n = 247$) are deferred at 13.8%; left-handed children ($n = 33$) at 10.9%. The differences are small.

6 Conclusion

We presented a normative-modeling framework for pediatric dysgraphia screening from digital handwriting features. The framework converts binary classification into a graded severity score

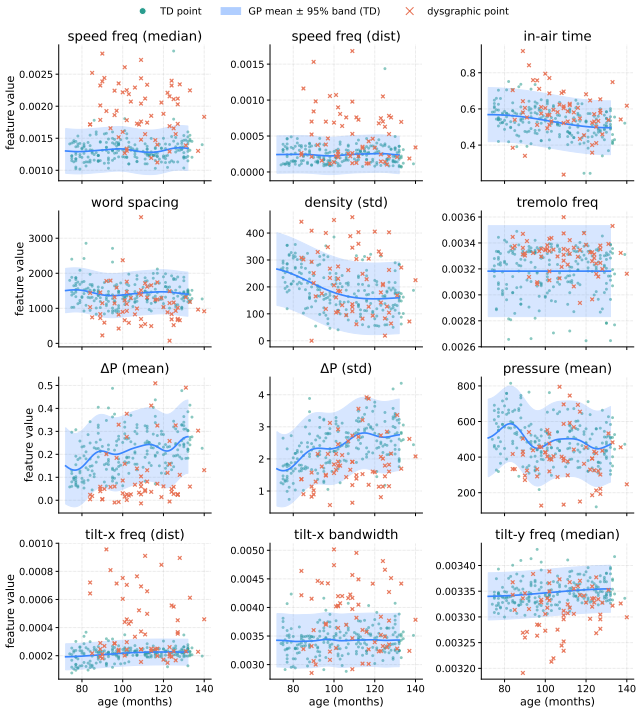


Figure 7: Per-feature Gaussian-process normative trajectories. Green dots are TD children used to fit the per-feature GPs; the blue curve is the GP posterior mean and the shaded band is the $\pm 1.96\,\sigma$ predictive interval. Red crosses are dysgraphic children, plotted at their chronological age. The 12 panels correspond to the 12 features used by the model.

(HMG) and an empirically calibrated triage system (Mondrian split-conformal). HMG carries information beyond chronological age, it captures a within-group severity gradient, and the conformal triage achieves the targeted empirical cross-validated coverage and singleton rate. Our framework provides a clinically interpretable score, an empirical uncertainty signal, and screening accuracy without a disadvantage relative to existing 12-feature classifiers. Our scheme offers a continuous, age-conditional representation suitable for finer-grained clinical decisions and longitudinal tracking. Future research directions include: external validation in independent datasets; extension of HMG to richer stroke-level or neural-network features; and longitudinal HMG trajectories for tracking treatment response.

Acknowledgments

The authors thank Christopher's Way Foundation for supporting this research and Thomas Gargot for providing the dataset.

References

[1] Anastasios N. Angelopoulos and Stephen Bates. 2021. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv:2107.07511* (2021).

[2] Thibault Asselborn, Mateo Chapatte, and Pierre Dillenbourg. 2020. Extending the Spectrum of Dysgraphia: A Data Driven Strategy to Estimate Handwriting Quality. *Scientific Reports* 10, 1 (2020), 3140.

[3] Thibault Asselborn, Thomas Gargot, Łukasz Kidziński, Wafa Johal, David Cohen, Caroline Jolly, and Pierre Dillenbourg. 2018. Automated Human-Level Diagnosis of Dysgraphia Using a Consumer Tablet. *npj Digital Medicine* 1, 1 (2018), 42.

[4] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.

[5] Maëlle Biotteau, Jérémy Danna, Éloïse Baudou, Frédéric Puyjarinet, Jean-Luc Velay, Jean-Michel Albaret, and Yves Chaix. 2019. Developmental Coordination Disorder and Dysgraphia: Signs and Symptoms, Diagnosis, and Rehabilitation. *Neuropsychiatric Disease and Treatment* 15 (2019), 1873–1885.

[6] Henrik Boström and Ulf Johansson. 2020. Mondrian Conformal Regressors. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications (Proceedings of Machine Learning Research, Vol. 128)*. PMLR, 114–133.

[7] Remco R. Bouckaert and Eibe Frank. 2004. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In *Advances in Knowledge Discovery and Data Mining (Lecture Notes in Computer Science, Vol. 3056)*. Springer, 3–12.

[8] Peter Drotar and Marek Dobes. 2020. Dysgraphia Detection Through Machine Learning. *Scientific Reports* 10, 1 (2020), 21541.

[9] Thomas Gargot, Thibault Asselborn, Hugues Pellerin, Ingrid Zammouri, Salvatore M. Anzalone, Laurence Casteran, Wafa Johal, Pierre Dillenbourg, David Cohen, and Caroline Jolly. 2020. Acquisition of Handwriting in Children With and Without Dysgraphia: A Computational Approach. *PLOS ONE* 15, 9 (2020), e0237575.

[10] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, Vol. 34. 18932–18943.

[11] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?. In *Advances in Neural Information Processing Systems*, Vol. 35. 507–520.

[12] Elisabeth Hamstra-Bletz, J De Bie, and Bernardus Petrus Leopold Maria Brinker. 1987. *Beknopte beoordelingsmethode voor kinderhandschriften: experimentele versie.* Swets & Zeitlinger. Concise Evaluation Scale for Children's Handwriting (BHK).

[13] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[14] Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. 2021. Second Opinion Needed: Communicating Uncertainty in Medical Machine Learning. *npj Digital Medicine* 4, 1 (2021), 4.

[15] Jayakanth Kunhoth, Moutaz Saleh, Somaya Al-Maadeed, Peter Drotar, and Younes Akbari. 2026. Grading of Developmental Dysgraphia Severity in Children: Multimodal Dataset and Classifier Fusion. *IEEE Transactions on Cognitive and Developmental Systems* 18, 2 (2026), 346–360.

[16] Charles Lu, Andréanne Lemay, Ken Chang, Katharina Hoebel, and Jayashree Kalpathy-Cramer. 2022. Fair Conformal Predictors for Applications in Medical Imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12008–12016.

[17] Andre F. Marquand, Iead Rezek, Jan Buitelaar, and Christian F. Beckmann. 2016. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological Psychiatry* 80, 7 (2016), 552–561.

[18] Jiri Mekyska, Marcos Faundez-Zanuy, Zdenek Mzourek, Zoltan Galaz, Zdenek Smekal, and Sara Rosenblum. 2017. Identification and Rating of Developmental Dysgraphia by Handwriting Analysis. *IEEE Transactions on Human-Machine Systems* 47, 2 (2017), 235–248.

[19] Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA.

[20] Saige Rutherford, Charlotte Fraza, Richard Dinga, Seyed Mostafa Kia, Thomas Wolfers, et al. 2022. Charting Brain Growth and Aging at High Spatial Precision. *eLife* 11 (2022), e72904.

[21] Xu Sun and Weichao Xu. 2014. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters* 21, 11 (2014), 1389–1393.

[22] Andrew J. Vickers and Elena B. Elkin. 2006. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making* 26, 6 (2006), 565–574.

[23] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer.

[24] Thomas Wolfers, Nhat Trung Doan, Tobias Kaufmann, Dag Alnaes, et al. 2018. Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA Psychiatry* 75, 11 (2018), 1146–1155.